STRATEGIES FOR EVALUATING THE IMPACT OF PROVINCE-WIDE TESTING

Principal Investigators
PHILIP NAGY
ROSS E. TRAUB

Research Officer
KATHERINE MacRURY

This research project was funded under contract by the Ministry of Education, Ontario.

It reflects the views of the authors and not necessarily those of the Ministry.

The Honourable Sean Conway, Minister
Bernard J. Shapiro, Deputy Minister

Ontario

STRATEGIES FOR EVALUATING THE IMPACT OF PROVINCE-WIDE TESTING

Principal Investigators
PHILIP NAGY
ROSS E. TRAUB

Research Officer
KATHERINE MacRURY

This research project was funded under contract by the Ministry of Education, Ontario.

It reflects the views of the authors and not necessarily those of the Ministry.

The Honourable Sean Conway, Minister
Bernard J. Shapiro, Deputy Minister

© QUEEN'S PRINTER FOR ONTARIO, 1986

Order Information:

Publications Sales
The Ontario Institute
for Studies in Education
252 Bloor Street West
Toronto, Ontario M5S 1V6
(416) 926-4707

Will invoice on orders over $30.00. Other orders must be accompanied by a cheque or money order payable to O.I.S.E.

Publications Services
880 Bay Street, 5th Floor
Toronto, Ontario M7A 1N8

(416) 965-6015
(Toll Free) 1-800-268-7540
(Toll Free from area code 807)
Ask operator for Zenith 67200
Order must be accompanied by a cheque or money order payable to the Treasurer of Ontario

---

ONO 3515

Canadian Cataloguing in Publication Data

Nagy, Philip, 1945-
  Strategies for evaluating the impact of province-wide testing

Bibliography: p.
ISBN 0-7729-1465-6

1. Educational tests and measurements--Ontario--Evaluation. 2. Grading and marking (Students)--Ontario--Evaluation. 3. Students--Ontario--Rating of--Evaluation. I. Traub, Ross E., 1938- . II. MacRury, Katherine A. III. Ontario. Ministry of Education. IV. Title.

LB3054.C36.N33 1986   373.12'62'09713   C86-099671-9

A GENERAL TABLE OF CONTENTS

ABSTRACT iv

PART ONE: Strategies for Evaluating the Impact of Examinations 1

1. Introduction and Background 4
2. Effects for the Examination Model 12
3. Studies that Draw on Ministry Data Bases 23
4. Surveys of Teachers 30
5. Studies of Perceptions 35
6. Controlled Studies of Teacher Marking Standards 38
7. Concluding Comments 40

PART TWO: Strategies for Evaluating the Impact of Assessments 45

1. Introduction 48
2. Possible Objectives of Assessment 52
3. The Facets 64
4. The Recommended Proposals 70

PART THREE: A Review of Literature on the Impact of Testing 75

A Preface to the Review of Literature 78
Section I: The Impact of Examinations 80
Section II: The Impact of Assessments 104

Appendices to Part Three

Appendix A: The Impact of Examinations,
A Selectively Annotated Bibliography 136

Appendix B: The Impact of Assessments,
A Selectively Annotated Bibliography 163

ABSTRACT

Part One: Strategies for Evaluating the Impact of Examinations

In this part of the report effects are identified that might be expected to occur in Ontario following the introduction of province-wide examinations for selected high school courses. Attention is focused on possible effects on

1. marking standards,
2. the curriculum,
3. enrolment trends, and
4. public perceptions of education.

To monitor these effects, four studies are proposed. The first would involve the analysis of data collected routinely by the Ministry of Education; this study would assess effects on enrolment trends and some aspects of effects on marking standards. A second study would be a survey of teachers to collect information about effects on curriculum, instruction and additional aspects of effects on marking standards. The third study would consist of public opinion polls to tap the effects of examinations on public perceptions. The fourth study would be an experiment on yet another aspect of the effects that examinations might have on marking standards.

If not all the proposed studies could be funded, the following factors should be considered in choosing among them: practicing educators, whether or not new data has to be collected, whether or not the study will provide a baseline for interpreting the information collected in subsequent studies, the extent to which the results are likely to be confounded by the effects of other changes in provincial education policy (e.g., funding of Catholic high schools), and cost. The key study is that addressing effects on the curriculum. Results from any of the other studies without information about what is taught and what is expected of students would be difficult to interpret.

Part Two: Strategies for Evaluating the Impact of Assessments

In this part of the report consideration was given to the effects that might be expected to follow the introduction of a provincial assessment of educational achievement, and studies were proposed for tracking those effects deemed most important. A case was made for the study of assessment effects

1. on variation in teacher marks,
2. on the evidence teachers collect and use to evaluate student achievement,
3. on the implemented curriculum,
4. on teacher marking standards,
5. on public perceptions of achievement in education and of its assessment,
6. on board policies governing the use of assessment results for evaluating personnel and for promoting professional development.

The second, third, fourth and fifth of these effects would be investigated by means of studies similar to those described in Part One of this report. Proposals are advanced in Part Two for survey studies that would address the first and last of the effects in the foregoing list.

If a choice had to be made among the proposed studies, attention should be paid to the matter of how the purpose of the assessment program might influence the effects that it could be expected to have. Also, as with the proposed studies of examination effects, a study of effects on the implemented curriculum seems essential for interpreting any other studies of the effects that an assessment program might have.

Part Three: A Review of Literature on the Impact of Testing

This part of the report is a review of the literature on testing effects. It is divided into two sections that deal, respectively, with examinations and assessments. The review was conducted with three objectives in mind:

1. the identification of testing programs in which the effects of testing have been monitored,
2. the identification of effects found to occur as a consequence of testing, and
3. the assessment of the applicability to Ontario of findings reported in the literature.

On the basis of the literature reviewed, it was concluded that studies of the effects of either an examination or an assessment program on Ontario education would not duplicate research done elsewhere. Previous studies were conducted on implemented testing programs, hence there was no opportunity to collect baseline data. In addition, the relevance of findings from other studies for Ontario is limited by differences between Ontario and the jurisdictions in which the studies were conducted.

PART ONE

Province-Wide Testing:

Strategies for Evaluating the Impact of Examinations

Part One

Table of Contents: The Impact of Examinations

1. INTRODUCTION AND BACKGROUND 4
    1.1. The Examination Model 5
    1.2. Context 6
        1.2.1. Province-wide testing since 1967 6
        1.2.2. The press for a return to province-wide testing 7
        1.2.3. The Ontario Government's response 8
        1.2.4. Confounding factors 8
    1.3. Activities of the Contract Team 9
        1.3.1. Literature search 10
        1.3.2. Study of data bases 10
    1.4. Overview of Remainder of Part One 11

2. EFFECTS FOR THE EXAMINATION MODEL 12
    2.1. Marking Standards 13
        2.1.1. Studying teacher marks 13
        2.1.2. Survey of teacher procedures for evaluating students ... 16
        2.1.3. Experimental investigations of marking standards 16
    2.2. Curriculum 17
    2.3. Enrolment Trends 18
        2.3.1. Confounding with OSIS 20
    2.4. Perceptions 21
    2.5. Advance Organizer 21

3. STUDIES THAT DRAW ON MINISTRY DATA BASES 23
    3.1. CROS File 23
        3.1.1. Limitations of the CROS File 24
    3.2. MR File 24
        3.2.1. Limitations of the MR File 25
    3.3. CE File 26
        3.3.1. Limitations of the CE File 26
    3.4. Studies of Marking Standards 26
        3.4.1. Studying marks without corollary information 26
        3.4.2. Studying marks with corollary information 28
    3.5. Studies of Enrolment Trends 28

4. SURVEYS OF TEACHERS 30
    4.1. Studying the Basis for Teacher Marks 30
    4.2. Studying Curriculum Effects 32

5. STUDIES OF PERCEPTIONS 35

6. CONTROLLED STUDIES OF TEACHER MARKING STANDARDS 38

7. CONCLUDING COMMENTS 40
   - 7.1. Comments on Priorities 41
   - 7.2. Limitations 42
   - 7.3. Closing Remark 43

LIST OF REFERENCES 44

CHAPTER 1

INTRODUCTION AND BACKGROUND

On 29 June 1984, the Ontario Ministry of Education sent out a call for proposals to evaluate the impact of province-wide testing. The overview to this call read as follows:

> The Ministry of Education is interested in the topic of Province-wide testing. Such a testing program might be helpful in addressing objectives such as the following:
>
> 1. reduce inequities to students caused by inconsistent practices in summative evaluation at the school level, and help post-secondary institutions and employers address the issue of mark variability;
>
> 2. address the needs of the public for assurance that standards of learning in our schools are recognized as important and are being maintained;
>
> 3. help meet the accountability and reporting responsibilities of the Minister to the public for the quality of education in Ontario;
>
> 4. help educators at the local level to compare the achievement of their pupils with provincial standards and to report results to parents and others.
>
> Province-wide testing would have to be compatible with societal values such as efficiency, co-operation and equity or fairness. Society would probably expect that these values be reflected not only in the tests themselves but in the overall experience of evaluation that students encounter in school.
>
> Ministry planners are faced with many decisions in attempting to develop a model of province-wide testing. The number of times a year that tests should be written, the number of subjects and grade levels for which tests should be developed initially and at later dates, the relationship between marks on province-wide tests and other measures of student achievement, the ways of recording and reporting marks, the relationship between province-wide tests and OAIP [Ontario Assessment Instrument Pool] pools, all are issues that could have far-reaching effects.
>
> The province-wide testing initiative is obviously an undertaking of such importance and complexity that there will be a need to monitor and evaluate its development and effects. Not only that, but it seems highly desirable that the monitoring and evaluation be planned at the same time that the model of province-wide testing and the instruments are being developed.

What follows is the report for a part of the study that was done under Ministry contract to identify the effects of province-wide testing and to design studies for monitoring and evaluating these effects.

When work began on this contract, the first order of business was to identify and describe in a general way the form of province-wide testing that would guide the work done on the project. This form is described in the next section.

1.1 The Examination Model

A province-wide testing program could emulate two general models:

1. the examination model

2. the assessment model

The examination model was chosen to guide the work reported here. (In a companion report, we consider the effects that might follow from adoption of the assessment model.)

As envisioned, the examination model would incorporate the following features:

- Examinations would be administered at the end of courses designated examination courses.

- Only secondary level courses would be designated examination courses, but these would not necessarily be courses taken during the last year of secondary school or courses leading to university entrance. (It is likely, however, that if examinations were introduced, there would be strong pressure to include end-of-school courses leading to university entrance among the examination courses.)

- An examination would be on the content -- knowledge, skills, understandings -- that constitutes the curriculum of the examination course.

- Every student in the province enrolled in an examination course would be tested.

- All students in the examination course would write the same examination; hence the performance of different students on the examination could be compared directly.

The examination model is severely limited in at least one respect:

> The number of questions that can be asked of each student is restricted to the number that students can be expected to answer in an examination period of reasonable length, say two or three hours. Thus, the depth and breadth of knowledge that can be tested is small in relation to the size of the curriculum.

The remainder of this introductory chapter is devoted to the accomplishment of the following objectives:

1. providing additional contextual information

2. listing the activities that were undertaken by the contract team

1.2 Context

1.2.1 Province-wide testing since 1967

Provincial examinations of achievement for secondary-school students were last administered in Ontario in 1967. That year and for several years thereafter, graduating students were eligible to take multiple-choice tests of scholastic achievement, English language proficiency, mathematics achievement and physics achievement. Scores on these tests, along with teacher marks, were made available to Ontario universities, to be used in deciding whom to admit to first-year programs.

The multiple-choice testing program was abandoned in 1974 because of the combined effects of two decisions:

1. The Ontario government discontinued funding for the development and administration of the tests.

2. Most Ontario universities decided not to ask applicants to submit test scores as a requirement of admission; hence there was no reason for students to write the tests and no possibility that the testing program could be funded from student fees.

Since 1974, no province-wide, every-pupil testing of any kind at any level has occurred in Ontario. University and college admission decisions are based on teacher marks and such other information as the schools and applicants supply.

This is not to say that the Ontario Ministry of Education has ignored student evaluation since 1974. As noted in the call for proposals that resulted in this project:

> In recent years, research on the testing of students in Ontario has centred around the development of the Ontario Assessment Instrument Pools. These open pools provide teachers and other evaluators with wide freedom of choice within the framework of the objectives in the provincial guidelines. None of these pools is complete and only two or three of them are at the OAC (university entrance) level. It is expected that more pools will be developed in the future, and that they will come to complement province-wide tests. The way in which they will in fact develop and complement one another, and the effects, remain to be determined.

1.2.2 The press for a return to province-wide testing

Pressure from a variety of sources has been exerted on the Ontario government for a return to some kind of province-wide testing. As summarized by Traub and McLean (1984), the following pressures have been applied:

- The Council of Ontario Universities has . . . been calling for a return to examinations in mathematics, English (for Anglophones) and français (for Francophones). [See Briefing Notes, No. 7. Toronto: Council of Ontario Universities, April 1984.]

- The Commission on the Future Development of the Universities of Ontario -- the Bovey Commission, as it [came] to be called -- [was] to "address a number of specific issues related to accessibility [to Ontario universities] such as the need for, and form of, general and specific entrance examinations to the Ontario university system, with reference to the new secondary school curriculum structure" (Minister's statement to the Legislature, 15 December 1983). [The eighth recommendation in the Commission's final report dealt with the matter of examinations as follows:

  > The Commission recommends that admissions direct from secondary schools be based on a combination of teachers' marks and school reports and of province-wide admissions examinations assessing achievement in at least language (English or français) and mathematics, but that alternative arrangements for admission of mature students be continued. (December, 1984. Ontario Universities: Options and Futures. Report of The Commission on the Future Development of the Universities of Ontario, P. 37.)]

- There is considerable editorial support for a return to provincial examinations. For example, in December 1983, editorials calling for school leaving examinations appeared in the Toronto Star [13 December] and the Toronto Globe and Mail [27 December]. Ostensibly these were in response to a report from Carleton University that most students admitted to the first-year program in computer science had failed mathematics. The editorial writers speculated that too many of the students admitted to Carleton's computer science program had been assigned marks in secondary school that were spuriously high compared to the marks assigned to other, more deserving, students, whose teachers had espoused higher standards of marking. In addition, the editorial writers expressed the opinion that "standards of education", whatever this phrase might mean, have fallen in Ontario since the discontinuation of the Grade 13 examinations. They seem to believe that school leaving examinations would reduce, if not eliminate, the inequities in university admissions due to variation in standards of marking from one secondary school to another and that such examinations would have the added effect of raising levels of achievement.

- Informed critics of education have called for a return to school leaving examinations. Mark Holmes, for example, has advanced the case for examinations (Globe and Mail, 3 May 1984; University of Toronto Bulletin, 22 May 1984), in part for reasons congruent with those advanced in the December 1983 Star and Globe and Mail editorials. Holmes also argued that examinations could be used to satisfy the public's need for accountability and to monitor changes in educational standards.

. Other jurisdictions have either returned to a system of examinations (e.g., British Columbia and Alberta) or never abandoned them (e.g., Newfoundland and Quebec). [Despite a common exam that counts for 50 percent of the final grade in Newfoundland, the teachers' components of the marks still vary considerably. (See Nagy, 1984.)]

1.2.3 The Ontario Government's response

On 20 March 1984, the Speech from the Throne that was delivered in the Legislature of Ontario contained the following announcement:

> In consultation with the Council of Ontario Universities and the Ontario Teachers' Federation, the Government will work to design a province-wide testing program necessary to assess the effectiveness of our curriculum and the performance of our students. The teacher in the classroom is the cornerstone of excellence in education, and, to a great extent, the promise of Ontario. However, to assist the Government in meeting its responsibilities, and parents in participating in their children's education, such tests will help all of us maintain the high quality of our educational system.

It was in this context that the Ministry of Education framed the call for proposals that resulted in this report.

1.2.4 Confounding factors

The form that a province-wide testing program would assume has yet to be determined. Whatever shape it might take, the effects it produces will be confounded by other significant changes now affecting Ontario's educational system:

1. A new secondary school program is being put into place; the policy statement in which this program is outlined is titled "Ontario Schools -- Intermediate and Senior" and is commonly referred to as OSIS. In this program, it is likely that some students, probably a small proportion, will find it possible to complete their schooling (not counting kindergarten) in 12 or 12.5 years instead of the present 13 years. Other changes, in the number of required credits and in the curriculum guidelines governing secondary school courses, are included in OSIS.

2. Roman Catholic Separate Schools, which formerly were publicly funded from Kindergarten through Grade 10, will soon be publicly funded for all grades. This change in funding is certain to mean that many students who would have gone to public secondary schools will go to separate schools. How this change will affect achievement and public perceptions of educational standards is not known.

3. Researchers who have been challenged to design ways to track the effects of a province-wide testing program have a further problem: Base-line data are needed so that in subsequent years it will be possible to compare the situation as it existed before the introduction of province-wide tests with the situation after their introduction. If province-wide tests are to be introduced soon, say in the 1985-86 school year, then the time available to collect base-line data is extremely short. If the tests are introduced in 1985-86, data collected during that year might approximate base-line data, depending on how the testing system is introduced.

It must be emphasized that the effects that will be associated with OSIS and the revised system for funding Catholic Schools are uncertain and subject to speculation, as indeed are the effects of province-wide testing. At the very least, however, we can expect changes of two kinds:

1. changes in the composition of the student bodies of many, if not most, Ontario secondary schools;

2. changes in the composition of the teaching force in many, if not most, schools.

Further, it can be said with assurance that the effects of OSIS and the new funding policy will combine with and otherwise confound the effects of province-wide tests, whatever form those tests are given.

1.3 Activities of the Contract Team

The objectives of this project were:

- to "review the literature";

- to develop "desirable and feasible options for a system or systems of monitoring [the effects of procedures] for evaluating ... student achievement in Ontario schools." (Quotations from the Request for Proposals for Research, p. 3.).

To achieve these objectives, the contract team undertook a number of activities, including reviewing the literature on testing effects, identifying the effects that should be monitored, and proposing studies of these effects. In addition, the contract team

- obtained information about the data bases maintained by the Ministry of Education, to see whether these might be made to yield useful information about testing effects;

- obtained information about data bases maintained by local school boards, to see whether or not such data bases might be useful; and

- held two meetings, one with experts in educational research and the other with a group of Ontario educators, to ascertain their views on the examination effects that should be tracked and the feasibility of mounting studies of these effects in Ontario.

Additional details are provided in the following two subsections of this chapter about the literature search and the study of data bases.

1.3.1 Literature search

- A search was made of the Education Index, which is a printed index of articles published in journals and educational newspapers, and of ERIC and Psychological Abstracts, which are computerized indexes. These searches considered the literature accumulated during the last 10 years under such keywords as achievement tests, assessments of education, examination, innovations in education, prognosis of student success, test bias, testing programs and test use.

- This literature was read, and sorted into various categories for summary and synthesis. Note was taken of the earlier literature referred to in these publications and the bibliography was increased.

- Letters asking for assistance in the identification of studies of examination effects were written to educational researchers in Australia, Great Britain, Israel, The Netherlands, New Zealand and Sweden. The publications identified in the replies to these letters were added to the bibliography.

- Several North American researchers were contacted by telephone, and asked to suggest publications.

- The literature identified in the various searches was summarized in an annotated bibliography. In addition, an analysis, evaluation and synthesis of this literature was prepared; it is contained in the third part of this report, to which the bibliographic references are appended.

The results of the literature search helped identify the effects to be studied and design the proposals that appear later in this report.

1.3.2 Study of data bases

Information was obtained about the data bases maintained by the Ministry and local school boards as follows:

- Meetings were held with the Ministry employees responsible for the Ministry data bases. In these meetings, the contents of three files, the CROS, MR and Course Enrolment files, were reviewed.

- Telephone contact was made with officials of Metro-Toronto and other Southern Ontario school boards to ascertain the nature and extent of the data bases maintained by local boards. It was judged that these data bases would not be useful in conducting studies of examination effects.

1.4 Overview of Remainder of Part One

The remainder of this report is devoted to developing proposals for studies whereby information can be obtained about four types of examination effects. The choice of effects was guided by reference to the examination model as described above. The important features of this model, which bear repeating here, are that every student in the province who is enrolled in the examination course would be tested, and the examination would be the same for every one of the students. In Chapter 2 we identify the four types of effects and provide a rationale for the decision to concentrate on only these types. The succeeding four chapters consist of proposals for four studies that could be conducted. Each proposal involves the study of an existing data base or the study of data that would have to be collected. Two of the studies would provide information about more than one type of examination effect. The seventh and final chapter of the report consists of a brief discussion of priorities.

CHAPTER 2:
EFFECTS FOR THE EXAMINATION MODEL

As a review of the contextual information presented in the previous chapter will reveal, examinations are offered by many Ontario educators and critics of education as a means of attaining several important and, to these individuals, desirable effects:

- raising standards of marking
- levelling standards of marking over teachers and schools
- improving the impressions that university faculty and admission officers have of the preparedness of students for university
- improving public perceptions of education

Other potential effects of examinations are mentioned in the literature or emerge from common-sense analyses. These include effects on:

- the methods teachers employ in evaluating students;
- the curriculum;
- enrolment trends.

An analysis of the foregoing list led us to group the effects into four categories:

1. marking standards
2. curriculum
3. enrolment trends
4. perceptions

Several other types of examination effects were discussed during the course of our enquiry. The foregoing categories exclude two effects of some significance:

1. No examination of the relation between university marks and high school marks has been recommended. Such a study would involve obtaining marks from the universities and matching them to information in Ministry files or the Ontario University Application Centre file; its cost would be high. Pooling data across universities or across

programs within universities would not be justifiable because of the differences in marking standards known to exist among universities and among programs within universities. (See Traub, Wolfe, Wolfe, Evans and Russell, 1976.) The sizes of the samples of students who obtained their secondary school marks from the same teachers and their university marks from the same professors would be very small; therefore this approach to studying differences in marking standards at either level -- secondary school or university -- would be unsatisfactory from a statistical point of view.

2. No investigation of changes in the racial composition of secondary school classes has been recommended. Collection of such information on a routine basis is unethical, if not illegal. Given the sensitive nature of the race issue, collection of such information in a research study is likely to prove difficult.

The remainder of this chapter presents our rationale for the types of examination effects that would be investigated if the studies proposed in subsequent chapters were actually conducted.

2.1 Marking Standards

As noted earlier, critics of education and the authors of newspaper editorials have voiced concern over declining and variable standards in the marking of educational achievement in Ontario. They believe that the mark assigned for a given demonstration of achievement has inflated since the Grade 13 departmental examinations were last administered in 1967. In addition, it is alleged that the mark one teacher awards is not the same as the mark another teacher awards for a comparable display of achievement in a course. In other words, there are soft as opposed to hard marking teachers. These allegations, if true, can be interpreted as evidence of declining and variable standards of marking.

2.1.1 Studying teacher marks

Beliefs about standards of marking are very difficult to support or discredit through the study of teacher marks. The main problem is that there exists no universally accepted measure of achievement for a course, a measure that might be used for the purpose of calibrating and comparing teacher marks. (A provincial examination is at best only one person's or one committee's approximation to such a standard.) Even if an absolute standard did exist it would soon become dated. The content of the curriculum changes over time, and an examination must pertain to the curriculum being followed, not one that is out-of-date. Moreover, examinations have a way of becoming public knowledge, even if they were to be kept secure. This is especially true of examinations that play a significant role in determining the futures of students. An examination in the public domain cannot be relied upon to differentiate students who have a general knowledge of the course from those who have only specific knowledge of the answers to the questions in the examination. Most examination programs for certifying educational achievement involve the production of new examinations for the next administration, even when the attempt is made to keep the examinations secure after they have been administered.

In the absence of universally accepted measures of achievement, attempts to understand and interpret the standards that underly teacher marks involve either accepting them at face value, invoking strong assumptions, or collecting corollary information.

Consider each of these approaches:

1. Accepting teacher marks at face value is difficult to justify because the assignment of marks is an exercise in relative, not absolute, judgment. To appreciate the standard implicit in a teacher's marks, it is necessary to appreciate the relational standard used by the teacher. It may be thought that this standard is the curriculum, and that a teacher's mark represents a corresponding (absolute) degree of mastery of the curriculum. This view cannot be taken seriously, however, for two reasons:

    a) The domain of knowledge for a course cannot be circumscribed precisely.

    b) This knowledge cannot be parcelled into units that are equivalent in any sense that is meaningful. Consequently, identical examination scores, which are earned through unequal performance on different items, do not necessarily represent the same level of achievement.

2. Consider, then, the use of assumption in the study of teacher marks. It might be assumed, for example, that the level of achievement reached by graduating secondary school students in a large population, such as Ontario's, is relatively constant from year to year. Were we to make this assumption, we might conclude that Ontario has suffered grade inflation and a decline in marking standards since the discontinuation of the departmental examinations. In 1968, the first year in which Grade 13 marks were determined without the benefit of an examination component, 69 percent of students earned an average mark of 60 or more for courses taken in their Grade 13 year. The percentages for the five years previous to 1968 varied from 55 to 47. It appears that marks inflated and marking standards dropped dramatically in the single year, June 1967 to June 1968.

    A different interpretation of these results is possible. There was a change from 1967 to 1968 in the way students were assessed. In 1968, the assessment would have been based only on the achievement a student's teacher expected to see displayed. This expectation stemmed from the teacher's detailed knowledge of what the student had been taught. The previous year, a student's mark would have been based only partially on the teacher's assessment; the rest would have depended on the student's performance on a departmental examination. Neither the student nor the teacher knew beforehand what would be tested by the examination. To the extent that the examination tested knowledge the student had not been taught, examination performance would have been below what it was on the teachers' assessments. By this line of argument, we see that marks may have been higher in 1968 simply because the assessment was based exclusively on what students had been taught, and that the marking standards used by teachers did not necessarily decline from one year to the next.

Trends in the average level of teacher marks since 1968 are not confounded by the change from a system with departmental examinations to one without; hence they might be more readily accepted as evidence of change in marking standards, at least by those prepared to assume no change in the true level of achievement of the population from one year to the next.

3. Rather than base interpretations of teacher marks on strong assumptions about reality, one can collect corollary information about student achievement and use this as the basis of an interpretation. Scores on an examination are one kind of corollary information. If we are prepared to accept an examination as the standard of achievement, then the marks different teachers award in a particular year can be studied in relation to examination scores for evidence of variation in marking standards from teacher to teacher; also the marks a teacher awards in different years can be studied for evidence of mark inflation and falling standards.

   The problems inherent in accepting an examination as the standard of educational achievement are serious. Not all well-intentioned educators can agree on the knowledge and intellectual skills that should be assessed by an examination. It is also obvious that some kinds of achievement cannot be assessed by written examinations. Many arbitrary decisions must be taken when an examination is prepared, administered and marked -- what to examine, what type of questions to use, how much credit to give each question, how much of this credit to give to particular responses, how long to let students write, what materials to let students refer to during the examination, and so on. The decisions that are taken will work in favour of some students and against others. To illustrate how this might happen, note that the students who write a provincial examination will differ to some extent in the opportunities they had to learn the knowledge and practice the skills required to answer an examination question. These differences will depend on the school attended (and within school, on the teacher), the textbooks read, the activities assigned for homework, and so forth. Thus, a student's score will depend on whether or not a given question is included on the examination. These issues stand apart from the well recognized problems of scaling that arise in comparing grades; these issues depend on a host of minute and discrete circumstances, which are peculiar to individual students and particular test questions.

This discussion suggests that the results of a study of teacher marks, with or without the corollary information contained in provincial examination marks, cannot be easily interpreted in terms of educational standards. Apart from fluctuation and inflation in actual grades, we might also ask about:

- differences among teachers in the kinds of achievements considered in the assignment of grades;

- differences among teachers in their perceptions of the quality of a display of achievement.

We discuss these differences in the next two subsections of the report.

2.1.2 Survey of teacher procedures for evaluating students

A study of marks is a sterile way to address the issue of educational standards. The substantive issue that should underlie any discussion of standards is what students know when they graduate from a course. Marks alone provide little information about standards in this sense. What would be informative is a study which identifies what evidence teachers collect and use as the basis of their assessments of student achievement. Class tests, end-of-term examinations, assignments, associated scoring guides, and so forth could be collected from different teachers of the same subject and compared in a cross-sectional study. This information would indicate whether one teacher's evidence differed from another's in the following respects:

- the form of the evidence -- whether one teacher relies more heavily on end-of-term examination results, another on class quizzes and assignments;
- breadth of curriculum coverage;
- depth of coverage;
- type of intellectual process that is tapped -- memory versus such higher order cognitive skills as problem solving and evaluation.

Information could also be obtained about whether one teacher differed from another in the weight given each kind of evidence. (e.g., one teacher might give class tests a weight of 50 percent in the final mark, whereas another teacher might give such tests a weight of only 25 percent.) A longitudinal design could be used to obtain evidence of change over time in the way students are assessed by their teachers.

2.1.3 Experimental investigations of marking standards

The study of teacher marks, however done, is inherently unsatisfactory; it affords no control over the stimuli to which the teachers (markers) respond. There exist real differences in the abilities and achievements of the students with whom different teachers work. Even if a common examination were administered to all students, differences in the abilities and achievements of the students in different classes would be reflected imperfectly in the examination scores, for reasons already discussed.

An alternative approach to the study of marking standards is the creation of an artificial situation in which teachers respond to a common set of materials. These materials might include "completed" examinations, project reports, essays and problem solutions. Teacher judgments of a common set of materials could be collected in a cross-sectional study and examined for variation in marking standards. If these judgments were collected in a longitudinal study, they could be examined for change in marking standards over time.

2.2 Curriculum

The decision to discontinue the provincial examinations in Ontario in 1967 was taken for several reasons, including the following:

- The examinations were said to be driving the curriculum in the sense that only that material tested by the examination was given much classroom emphasis. The perception that "teaching to the test" was a problem probably varied from one subject to another because achievement in some subjects can be assessed more adequately by means of a written test than can achievement in other subjects.

- Some educators took a dim view of the industry that developed around the provincial examinations. For example, Coles' Notes were developed as study aids. They contained reprints of old exams and compendia of the information needed to score well on the provincial exams. The production of these notes and the appearance of schools for coaching students in the examinations were seen as proof by some observers that factual recall was emphasized to an undesirable extent by the examinations and that cramming was endorsed as a valid way of preparing for the examinations. (See Brown, 1967a, 1967b.)

- It was a widespread practice of teachers to cover by the early spring all the curriculum that would be examined. The last few weeks of the academic year were then available for preparing the students for the exams.

- Teachers believed that the quality of their teaching was assessed by their success in getting students to pass the departmental exams.

On the basis of these arguments, it was widely believed that the Grade 13 examinations influenced the curriculum in undesirable ways. Some of the rhetoric encountered in the literature review is in line with these beliefs; it argues in favour of studies to see whether an examination system would indeed have some of the aforementioned consequences.

There is, perhaps, a case to be made that at least one effect of examinations on the curriculum would be favourable: if the pressure of an external exam were to prevent some teachers from deviating from the curriculum guideline, then a province-wide examination might be said to have had a beneficial effect. This would almost certainly be the view of those college and university professors who want to assume that all first-year college and university students possess a common background in, for example, mathematics. Against this however, are two arguments:

1. that every curriculum contains some topics and skills, the achievement of which cannot be evaluated by written examination;

2. that teaching (and, in consequence, learning) is often best when the topic is one in which the teacher is vitally interested, no matter how idiosyncratic the topic.

The danger is that these topics and skills would be de-emphasized by teachers under the pressure of an external exam. This is but another facet of teaching to the test.

From this it is clear that a number of questions would arise if province-wide examinations were reintroduced. These questions include the following:

. What emphasis is currently placed on particular course objectives? How much does this emphasis vary from teacher to teacher?

. Would the introduction of exams cause teachers to concentrate on a narrower set of objectives, in particular those objectives that are amenable to assessment by written exam?

. Would teaching to exams, practice on old exams, and special coaching (cramming) materials become commonplace following the introduction of exams?

. Would the form of instruction be tailored to the type of knowledge (e.g., factual recall) required for scoring well on written exams?

. Would these effects filter down to the earlier grades?

These questions could probably be answered best by observing teachers at work in their classrooms over an extended period of time. The cost of large-scale observation studies is prohibitive. An approximation to the ideal study could probably be achieved by asking teachers to report on their practices. Surveying teachers by questionnaire and having teachers log what is taught, when and how, are examples of procedures that could determine the impact of provincial examinations on the curriculum. Teachers could be asked to record the emphasis (or time spent) on different aspects of the curriculum and in different instructional modes. Data collected now would establish a baseline for the situation before examinations had been introduced; data collected after would make it possible to track the effects of examinations.

2.3 Enrolment Trends

The call for proposals that resulted in this report suggested that the effects of province-wide testing on student enrolments and minority groups should be studied. A related issue, which arose as the report was being developed, is the effect provincial examinations would have on the pattern of courses selected by students during their secondary school years. Effects such as these would be addressed in studies of enrolment trends.

It seems obvious that it is important to consider how many students and which kinds of students are taking particular secondary school courses; these data provide a means of interpreting the evidence collected about achievement by the administration of examinations. It is well known that the scholastic achievements of secondary school students in the United States, as reflected by several different indicators, including scores on the Scholastic Aptitude Test, declined during the decade from 1965 to 1975. How to explain this decline, which recent evidence suggests has been arrested, was the subject of considerable speculation and debate for a time. An appealing hypothesis, in keeping with the present trend toward a more traditional view of education, is that secondary school students were opting, during this decade, for less challenging courses than students had selected in previous decades and that students may now again be choosing. If information had been available on the courses that students had taken, this hypothesis could have been put to the test. The tracking of students' programs would permit the Ministry of Education to review from time to time its policies on required courses.

Provincial examinations would be a powerful mechanism for shaping the behaviour of students. It seems almost certain in light of the Council of Ontario Universities' expressed wish for examinations -- see Chapter 1 -- that Ontario universities would require students to take one examination or more as a condition of admission, especially if several Ontario Academic Courses (OACs) had provincial examinations. (Whether more than one score would be required would likely depend on the courses for which examinations were developed and on the university program to which application for admissions had been made.) If the role of examination scores in admissions to post-secondary institutions (e.g., university, college of applied arts and technology) were significant, students would want to achieve as high an examination mark as possible. Enrolments might be affected as follows:

- Students would try to progress quickly through the sequence of courses leading to the courses with examinations, to be in a position to rewrite an examination if the mark on the first attempt were low. An implication of this line of thinking is that students would enrol in the examination courses earlier in their secondary school careers than would be the case if these courses were taken in "normal" sequence. Enrolment trends could be checked to see whether or not the students taking the courses in a sequence leading to an examination were younger than the students in a sequence of courses not leading to an examination.

- Another consequence is that summer school enrolments in those courses in sequences leading to examinations would increase.

- More students would be motivated to take the courses with examinations that was previously the case. Enrolments in optional courses (i.e., those not required under OSIS) might fall off.

- In semestered schools, it would become accepted practice for students to take examination courses in the first semester, so that their examination scores would be available when they applied for September admission to university. (These applications are due at the Ontario University Application Centre in Guelph in April, long before students in second semester courses would have written the examinations -- late May or early June.)

- For the same reason, there would be an increase in the number of semestered schools.

- If the provincial examinations were in OAC courses, more students than at present would decide early in their secondary school careers to try the OAC (advanced) course stream. This would allow them to keep open as long as possible the option of post-secondary study. In later grades, then, there would be a larger number of transfers than is now the case from the OAC stream to the general course stream.

- There would be an increase in the number of students who apply to post-secondary institutions without the examination qualification.

One can speculate too about the effects of provincial examinations on the pattern of courses taken by various subpopulations of college and university bound students. Would examinations attract or repel students who are members of racial minorities? Would students from homes in the lower social classes be attracted or repelled by the examinations? Would these students find it easier or more difficult to gain access to college and university? Would an examination in mathematics result in fewer female students taking courses in this subject? All of these questions seem *a priori* to be of interest, and could be answered by a comprehensive study of enrolment trends. (We note here the fact that existing data bases do not contain the information needed to answer all these questions. Moreover, we do not propose studies in which the additional data needed to answer all these questions would be collected. It may be desirable for the Ministry to consider expanding the data bases it maintains, to include information on courses taken all through secondary school and information on additional demographic characteristics of students.)

2.3.1 Confounding with OSIS

We must be careful not to expect more from a study of enrolment trends than such a study could deliver. The present report was motivated by the challenge to design studies of the effects of province-wide testing. As has been noted, other significant factors in addition to the introduction of examinations are impinging on the education system in Ontario. In particular, the introduction of OSIS itself will have a profound effect on enrolment trends. For example, a student who decides to take the requirements for university entrance in four instead of five years will be limited to a relatively narrow set of course selections. Such a student will face a large set of compulsory courses (16), and these, along with the sequenced courses that are required for the optional subjects that are taken, will leave little freedom of choice in a four-year program. Substantial shifts in enrolment patterns might therefore occur for reasons unrelated to the introduction of provincial exams.

2.4 Perceptions

According to the terms of reference of this project, public and university perceptions of the quality of secondary school graduates should receive careful attention. It is worth noting that most public criticism of education has been advanced without the support of systematically collected data. The importance of responding to critics by referring to this kind of data is itself reason enough to carry out the proposed projects.

In much of the public debate over standards in education, the separate issues of grade inflation (over time) and mark variation (between schools) have been intertwined. These two issues probably differ in the importance that different audiences would place on them. It seems likely that members of the public are concerned about grade inflation, given the widely held perception that large numbers of high school students lack basic skills when they graduate. On the other hand, university faculty are probably more concerned about variation in the backgrounds of the students they must teach. Although this point is open to empirical investigation and to discussion, it serves to illustrate the nature of the information that should be gathered about perceptions.

The manner of collecting information about perceptions is straightforward enough: sample surveys of selected audiences, such as Gallup Polls. To track changes in perceptions over time, information would have to be collected on a regular basis, perhaps every other year. To compare the opinions of different groups, each of which may have more valid opinions on some issues than on others, various subpopulations would have to be considered separately. Six subpopulations of possible interest are as follows:

- students
- parents
- teachers
- school administrators
- faculty of post-secondary institutions
- admissions officers of post-secondary institutions

2.5 Advance Organizer

This completes the rationale. After we had considered how to assess or track effects of the four types discussed in this chapter, we decided to organize the remainder of this report, not according to the four types of effects, but according to the different data bases that would be studied. This approach provides for more efficient use of resources: at least two of the proposed studies would address more than one type of effect. We turn next to a consideration of proposals for specific studies.

(Several of the proposed studies, in modified form, are suggested in the second part of the report on the potential effects of the assessment model.)

CHAPTER 3:
STUDIES THAT DRAW ON MINISTRY DATA BASES

The data bases maintained by the Ministry of Education could yield information on at least two kinds of effects: marking standards and enrolment trends. As noted earlier, these data files are known as the CROS, MR and Course Enrolment (CE) Files. If province-wide examinations were introduced, another file would probably be developed and maintained, the Examination Score (ES) File. Before we consider the kinds of studies that could be mounted using these files, it is necessary to be more specific about the contents of the CROS, MR, and CE Files. Should examinations be introduced, the contents of the ES File are easy enough to imagine.

3.1 CROS File

This file contains information about each student enrolled in at least one Secondary School Honour Graduation Diploma (SSHGD) course during the school year in which the file is compiled. (A new version of the file is compiled each year.) Obviously, the data in the CROS File can provide information only about OAC courses, and only then if the information currently gathered for SSHGD courses will also be gathered for OAC courses.

The information in the CROS File includes the following:

- Name, address, birthdate, sex, Ministry Identification Number, and Ontario University Applications Centre (OUAC) Number, if the student has one;
- Student's status in Canada (citizen, landed immigrant, student visa, other);
- Marital status;
- Language first spoken and still understood;
- Expectation for fulfilling the requirements of the Secondary School Honour Graduation Diploma (SSHGD) by the following June;
- Grade (11, 12, or 13), and if 13, whether it is being repeated or not;
- Whether the student will be taking Grade 13 next year, or intends to take a full time job or apply for admission to a university, Ryerson Polytechnical Institute or a College of Applied Arts and Technology;
- Whether or not the SSHGD was recommended and awarded;
- Whether or not the student was awarded an Ontario Scholarship;
- Whether or not the student applied for admission to university;

- For each SSHGD course taken and passed, the course code, course credit value, and student's mark; (Schools are not allowed to report the failing marks of students in any honour graduation course.)

- The kind of school attended, whether day or night school.

3.1.1 Limitations of the CROS File

Despite the considerable amount of data contained in the CROS File, it is limited in several important respects:

- There are only six years of data in the computer-readable bank of CROS File data: the 1983-84 file is in the process of being completed.

- The Social Insurance Number was used as the student identifier in the CROS Files for 1977-78, 78-79, 79-80 and 80-81. Since then, an unrelated Ministry Identification Number has been used. The records of students who appear in one or more of the files for 1980-81 or earlier and also in a later file cannot be connected.

- Not all the information on students who appear in more than one CROS File is "rolled over". Specifically, until the present year (1983-84) only marks of 65 or more in SSHGD courses were carried over to the next file. This means that if complete SSHGD information were required, it would be necessary to search the files of previous years, where possible, in order to update the records of those students who took SSHGD courses in more than one year.

- Whenever marks are carried over, no indication is given of whether or not the previous marks were obtained in the same school.

- It is not possible to relate a student's marks in a course to a particular teacher unless the student attended a school with only one teacher for the course. Similarly, it is not possible to relate a student's marks to a particular class in the course unless the student attended a school with only one class in a particular subject.

- Failing marks are not submitted, hence are not recorded. It is not possible to obtain information from the CROS file about failure rates.

- Perhaps the most serious limitation to the CROS File, at least for the purpose of studying enrolment trends, is the fact that it contains information about SSHGD courses only.

3.2 MR File

The MR file is compiled from the Secondary School September Report. It includes the following information:

- Data on total student enrolment, with the numbers of students who transferred to other Ontario schools, retired from school, or were admitted during the previous year;

- A variety of other information on students --
  - honor graduation credits taken by students other than grade 13 students, cross tabulated by number of credits and grade;
  - summer school credit courses cross tabulated by language of instruction and by level of course within the categories of newly attempted and repeated courses;
  - cross tabulations of numbers of students by age and sex, and by grade and sex;
  - numbers of students who expect to receive graduation diplomas of the three different kinds -- Secondary School Graduation Diploma, Secondary School Honour Graduation Diploma and Certificate of Training --in January and in June;
  - enrolments in courses in the official language (English or French) of students for whom that language is not their first language;

- Data on individual teachers:
  - duty (e.g., principal, department head, regular teacher);
  - assignment to special programs (e.g., library, guidance, special education), if any;
  - years of experience by level (elementary, secondary, other);
  - language of instruction (English, French);
  - grades taught;
  - subject areas taught (including the ambiguous designation "multi-subjects").

3.2.1 Limitations of the MR File

The data on students are limited in that they are aggregated to the school level. Thus, the information in this file cannot be related to the information on individual students in the CROS File.

The data on individual teachers are limited in that they do not indicate special certificates or special areas of training. This information would be of interest because one possible consequence of shifts in enrolment patterns might be a change in the qualifications of those teaching a particular course.

3.3 CE File

This file is assembled from the School September Report on Course Enrolment. It includes the following information on each course, whether that course is based on a Ministry Guideline or not:

- Level of the course -- whether intermediate, senior, or honor graduation, and within intermediate, whether basic, general or advanced.
- Number of classes per course.
- Credit value of the course.
- Total enrolment in the course by sex.
- Where appropriate, information is given separately by language of instruction, whether English or French.

3.3.1 Limitations of the CE File

Since this file is based on a September Report, it does not contain information about subsequent changes: drop-outs, transfers into courses, and so forth. But this file, together with the CROS file, could yield information on SSHGD or OAC dropout rates.

3.4 Studies of Marking Standards

3.4.1 Studying marks without corollary information

An investigation of marks could be made in which only the CROS File is used. This study would be directed to questions such as the following:

1. Is there evidence of province-wide inflation or deflation in teacher marks since 1977-78?

2. Is there evidence of variation from school to school in extent of mark inflation or deflation since 1977-78?

3.4.1.1 Unit of analysis

Because an individual teacher is responsible for collecting the evidence on which a student's mark in a course is based, the teacher should be the unit of analysis in any study of marks. As stated above, an analysis of the marks contained in the CROS Files by teacher or class is not possible. The analysis of teacher marks can be no more fine-grained than the aggregation of student data to the level of the school. (Analyses at the class or teacher level are not possible.) Even then, marks will probably be encountered that cannot be referred clearly to a school. Still, it should be possible to relate the marks of the vast majority of students in the CROS File to a school, and thus obtain a distribution of marks within a school for a course. This will make it possible to follow changes through time in mark distributions for a school, as well as for the province as a whole.

3.4.1.2 Research for the Bovey Commission

A study of marks was undertaken by the Ministry of Education for The Commission on the Future Development of the Universities of Ontario. The Ministry compiled distributions of marks for each year from 1978 to 1983 for the province as a whole. Results for different courses were pooled to obtain one distribution for each year. What is proposed here would not duplicate that effort. A school-level analysis is needed to discover how widespread grade inflation has been in recent years, and to provide baseline data for future reference.

3.4.1.3 Proposal

A study of marks at the school level would be conducted roughly as follows:

- Copies of CROS Files would be obtained for the three years -- 1977-78, 1980-81, and 1983-84.
- A decision would be taken on which course marks would be investigated. Prime candidates would be those courses with large numbers of students and standard Ministry Curriculum Guidelines or Courses of Study.
- The marks earned by the students who took the chosen courses in one of the three years would be drawn from the CROS Files.
- The files of marks would be processed to obtain the mean mark in a course for each school in each year under study, and also to obtain within-school and between-school indices of mark variation for each year.
- For each course, mean marks (with an indication of variance) for the three years would be plotted for selected schools (e.g., for schools at the extremes and at the 10th, 25th, 50th, 75th and 90th percentiles of the distribution over schools of the difference between the mean mark for 1977-78 and that for 1983-84).

This analysis could provide information on the extent of grade-inflation for the time period 1977 to 1984. Moreover, the phenomenon of grade-inflation could be linked to specific schools, so that associations between extent of inflation and demographic characteristics of schools could be considered. The main difficulty with this analysis is that without corollary information, what we are calling grade-inflation could not be separated from real changes in the quality of the student cohort within a school.

3.4.2 Studying marks with corollary information

If a provincial examination system were introduced for (at least some) OAC courses and if the Ministry were to compile CROS Files year-by-year for OAC-level courses as is now done for SSHGD-level courses, then examination scores would be available for the students taking the OACs with examinations. And if the examination scores were stored in a computer file (ES File) with students identified by their Ministry Identification Numbers (MINs), then it would be relatively easy to link a student's teacher-assigned marks in the CROS File to his or her examination scores in the ES File. This would make it possible to study the relationship between teacher-assigned marks and examination scores, to see whether, in a cross-sectional study, this relationship varies substantially from school to school. (In a study conducted two or more years after the introduction of provincial examinations, it would be possible to compare the teacher-assigned-mark by examination score relationship observed for a school in one year with the relationship observed for that school in the next year.)

3.4.2.1 Proposal

In the first year of province-wide examinations, a cross-sectional study of the teacher-assigned-mark by examination score relationship would be conducted as follows:

- A copy of the CROS File would be obtained for the year.
- A copy of the file of examination scores would be obtained.
- Marks and scores for those courses with examinations would be matched using MIN codes.
- A variety of exploratory analyses would be undertaken, beginning with plots of marks versus examination scores and following with such statistical analyses as seem reasonable.

These analyses would obtain evidence of inter-school differences in standards of marking and provide baseline information for future reference. The study would measure the extent of variation over schools in the strength of the relationship (correlation) between teacher marks and examination marks. The extent of variation among schools in the average of the teacher assigned marks for students who obtained a particular examination score could also be studied. The differences among schools that contribute to these kinds of variation could be related to demographic variables.

3.5 Studies of Enrolment Trends

Straightforward enrolment counts based on the data in Ministry maintained files could be used to answer several interesting and important questions of a longitudinal nature, provided the information collected by the Ministry did not change over the period of time under study.

The Ministry files would be useful for tracking changes in course enrolments. For example, the MR file could be used to report on:

- changes in the percentage of students in courses with provincial examinations;
- changes in enrolments in summer school courses;
- changes in numbers of January graduations;
- changes in numbers of students taking French as a second language.

From the CE File it would be possible:

- to track changes in enrolments in the sequences of courses that lead to provincial examinations, and compare them to changes in enrolments in the sequences of courses that do not end in a provincial examination;
- to track changes in enrolments in courses by sex, perhaps cross-tabulated by language of instruction.

CHAPTER 4:
SURVEYS OF TEACHERS

The limitations of the data bases maintained by the Ministry for answering questions about examination effects were discussed in the previous chapter: if more than a cursory study of examination effects were to be made, additional data would have to be collected.

In this chapter, we consider the data that could be collected by surveying teachers. The surveys proposed here would involve:

- interviewing a relatively small number of teachers;
- sending questionnaires to a relatively large number of teachers;
- asking a sample of intermediate size to supply materials of some kind;
- having a small number of teachers maintain logs of their teaching activities.

Data of this kind could be used to answer questions about the effect of province-wide examinations on marking standards and the curriculum. Data could also be collected in these ways to answer questions about teacher perceptions, but these questions are dealt with separately in the next chapter.

4.1 Studying the Basis for Teacher Marks

The possibility that different teachers assign different marks for achievement at the same level is generally viewed as a matter of concern. A study of marks alone could not reveal whether or not differences in the marks assigned by teachers were associated with real differences in student achievement. The lack of a fixed and absolute scale for measuring achievement and comparing the marks of different teachers is the Achilles heel of any study of marking standards. Our only recourse would be to collect data that would inform us about other differences among teachers, differences that might be related to differences in marks.

One of the other ways in which teachers might differ is in the evidence they choose to collect and use in assigning marks. For example, one teacher might rely heavily on class tests and term examinations in assigning marks in mathematics, whereas another teacher might rely much more on performance of take-home assignments. Also, different teachers might emphasize different areas of content and different cognitive abilities (e.g., recall versus higher-order abilities) in their assignments, class tests, etc. Such differences do not translate into simple assertions about differences among teachers in the way they associate mark to quality of student achievement. Differences in the relationship between marks and quality of achievement may well exist, but in addition there may exist qualitative differences among teachers in the manifestations of achievement that are considered. At this time little is known about such differences. A study has been made of the examinations used by Ontario teachers of English at the SSHGD level (Graham, 1984). No other systematic study has been made recently of the student evaluation procedures that Ontario teachers employ. The following questions are important to this issue:

- On what basis are marks now awarded? What proportion of marks are assigned for written tests, projects, essays, etc.?
- How does this basis vary across teachers?
- What contents and types of cognitive skills are assessed? Do these differ substantially from teacher to teacher?
- Would the introduction of examinations be associated with a substantial change in the evidence used by a teacher in assigning marks, and in the way that evidence varies from teacher to teacher?

4.1.0.1 Proposal

To obtain answers to the above questions, a sample of teachers could be surveyed as follows:

- Interview a few -- say 10 -- teachers of a subject, finding out what evidence of academic achievement is collected; obtain copies of examinations, assignments, quizzes, and the like.
- Use the interview information to design a questionnaire to elicit similar information from a large sample of teachers.
- Test the questionnaire on a few -- say 5 -- teachers and revise it on the basis of test results. (This step may have to be repeated more than once.)
- Send the questionnaire to a randomly selected provincial sample of 100 teachers.
- Have the teachers complete the questionnaire and ask each teacher to submit copies of the quizzes, assignments, examinations and so forth that he/she uses in assessing student achievement in the course.
- Tabulate the questionnaire responses, and analyze them in relation to the teacher's training and experience.
- Analyze and compare the materials submitted by the teachers, again in relation to the teacher's training and experience.

- Interview a subsample of these teachers to probe responses and validate the questionnaire data.

This type of survey should be conducted for each subject with provincial examinations and for at least two subjects without examinations. This would make it possible to detect any differences between examination and non-examination subjects in teachers' reliance on examinations as distinct from other devices for evaluating student achievement; it would also be possible to discover differences in emphasis on memorizing facts as opposed to higher cognitive skills (e.g., ability to apply knowledge, ability to solve problems, ability to evaluate).

[At some point, not necessarily the first year after examinations were introduced, it would make sense to survey teachers in pre-examination grades who teach courses in the sequence leading to an examination course. The survey of these teachers might involve only a questionnaire (perhaps that administered to teachers of the examination course, but in revised form). This step would add little to the cost of a study, and would provide a basis for assessing the extent to which exam effects had filtered down to lower grades.]

Further on this proposal:

Comparing results across curriculum areas would be difficult because the traditional methods of student evaluation in different subjects are themselves different. Also, achievement in some subjects is more amenable to assessment by examination than is achievement in other subjects.

Longitudinal comparisons within a subject might detect change over time in the way students are evaluated by their teachers. Baseline data would result from this study. It could be used in a future study as a basis of comparison, to see whether there had been a change in the evaluation procedures used in a subject, and to see whether the changes were more dramatic in examination than non-examination subjects.

A beneficial side effect is that the study would provide information for planning in-service programs for teachers on the topic of student evaluation.

4.2 Studying Curriculum Effects

As experience in Ontario before 1967 has shown, and as information from other jurisdictions confirms, there is good reason to expect that provincial examinations would have a substantial effect on curriculum and on associated teaching practices. With examinations, teachers could be expected to emphasize that material which is examined and to downplay that material which is not. Thus, information should be collected about the emphasis teachers put on particular objectives, and about how this emphasis changes with the introduction of exams.

Among the issues that should be addressed in a study of the effects of province-wide examinations on the curriculum are the following:

- the extent to which there are differences among teachers in the objectives they espouse, both in the absence of provincial examinations and in their presence;
- the extent to which there is a narrowing of course objectives following the introduction of provincial examinations;
- the extent to which there occurs teaching to the test, both in content and in focus on types of items which can require different kinds of cognitive function (e.g., recall versus higher-level cognitive skill);
- the extent to which use is made of old exams and special coaching tools.

There are two reasons for recommending the collection and analysis of this information:

1. The information itself is important because it would indicate whether or not an effect had occurred; debate could then focus on the issue of whether that effect were beneficial or harmful. (Presently, this debate occurs without knowing whether or not there is anything to debate.)

2. The information would provide a context for interpreting the findings obtained in studies of other examination effects, particularly those involving the general issue of standards.

Also, side benefits, such as those arising from information collected in the study of evaluation practices, could be expected. The information collected about classroom practices would be valuable to those who study curriculum implementation and, if collected over time, to those interested in curriculum change.

4.2.0.1 Proposal

The following steps parallel those suggested for the study of the bases teachers use in assigning marks:

- A small sample of teachers would be interviewed to ascertain course content and teaching practices.
- With the assistance of subject matter experts, the data from these interviews would be used to devise questionnaires and log-books. The questionnaires would collect information on teacher perceptions of their curricular emphases. The purpose of the logbooks would be to help teachers record, over an extended periof of time, information about their teaching practices and curricular emphases. The questionnaires and logbooks would be tested on a small sample of teachers, and revised. (This step would probably require repetition.)

- The questionnaires would be administered to a random sample of 100 teachers; the logbooks would be administered to a sample of 25 teachers. (An honorarium would probably be required to induce teachers to take on the onerous task of keeping a logbook.)

- The data analysis would be designed to portray curricular practices in a simple descriptive fashion.

To be most useful, this study should be done at least three times; the first study would be conducted before the introduction of examinations. The study should then be repeated two years later, and again two years after that. This cycle should be completed for each subject in which there were a provincial exam, and for at least two subjects in which there were no exam. This would make it possible to compare what happened in examination subjects and in non-examination subjects.

CHAPTER 5:
STUDIES OF PERCEPTIONS

The effect of provincial examinations on perceptions would be easy to study compared to the types of effects dealt with in the previous chapter. A survey questionnaire would have to be developed, with questions addressed to the examination issues of interest. (e.g., Do you think the introduction of provincial examinations would improve the standard of education? And after the examinations have been in place for a time: Now that we have province-wide examinations, are graduating students better prepared for university than they were before?) The questionnaire would be administered to a random sample of the target population of respondents.

Several subpopulations might merit special attention:

1. students, specifically
   a. those in high school,
   b. those in first year of university,
   c. those in first year of college;
2. members of the public subdivided on the basis of whether or not they have children, and if so, whether they have a child in secondary school, college or university;
3. secondary school teachers;
4. school administrators;
5. school trustees;
6. post-secondary faculty, subdivided by institution -- university versus college;
7. post-secondary admission officers, also subdivided by type of institution;
8. business and community leaders, including the personnel officers of large companies;
9. members of the provincial legislature.

5.0.0.1 Proposal

The basic procedure for a study of perceptions could be as follows:

- Interview a few members of the subpopulation.
- On the basis of interview data, design questionnaires for the subpopulation.
- Test the questionnaires and revise them. This step might have to be repeated, perhaps more than once.
- Hire a professional polling company to administer the questionnaires to a random sample of the subpopulation.
- Tabulate and analyze the responses following accepted survey methodology.
- For the second and subsequent surveys, compare the results with those from earlier surveys.
- Prepare a written report of the results.

We do not list here the questions that should be asked in the questionnaire. These would vary depending on whether the population at large was being surveyed or some subpopulation. The intent of the questions would be to examine perceptions of:

- examination effects on standards;
- examination effects on the quality of secondary school graduates;
- examination effects on admission decisions to post-secondary institutions. (Are these decisions seen to be made with greater fairness after examinations have been introduced than before?)

There would also be questions pertinent to certain subpopulations:

- For students, questions about:
    - student attitude;
    - teacher attitude;
    - quality of preparation for exams;

- role of the teacher.

. For teachers, questions about:

  - student attitude;
  - teacher attitude;
  - exams as teacher assessment devices.

CHAPTER 6:
CONTROLLED STUDIES OF TEACHER MARKING STANDARDS

The issue of marking standards is multifaceted. In Chapter 3 we proposed studies of teacher marks and examination scores that would yield results bearing on standards. In Chapter 4 we proposed surveys of teachers to ascertain the extent to which there are differences among them in the kinds of evidence collected and used when marks are assigned.

The standards teachers apply in marking examinations, essays and other evidence of academic achievement are another aspect of the standards issue. If examinations reduce variance in the meaning of marks assigned by different teachers or schools, this effect will be partly due to the fact that teachers across the province acquire a more uniform shared understanding of the standard of performance to be expected for academic work in a course than they had in the absence of examinations. We assume that Ontario would follow the practice of other provinces in which provincial examinations are administered and make the examinations public after they had been given. It therefore seems reasonable to suggest that greater uniformity of marking standards should develop after the first provincial examination for a course has been administered and made public; teachers would then be able to see what the examination developers expect students to achieve.

To study differences among teachers in the standards they apply, a sample of teachers would be asked to respond to a common set of stimulus materials. These materials might consist of class tests, end-of-term examinations, term papers and so on. This kind of control on marking is needed to avoid two effects that may to confound the interpretation of differences among teachers in the marks they assign:

1. The materials would represent a common set of students; consequently, no real differences in student achievement would underly the judgments made by the different teachers.

2. Since the common set of materials would not be associated in the teachers' minds with real students, no halo effect would contaminate their judgments.

Because marking would be under the control of the researcher, it would also be possible to introduce ancillary information (e.g., information about the extent to which the teacher's mean mark departs from the mean of all teachers' marks) into the marking situation. This could be done to determine the effect of the information on the marks that were assigned.

A limitation of the proposed study is that the absence of personal information on students would reduce the match between the conditions of the study and real teaching conditions.

6.0.0.1 Proposal

A controlled study of teachers' marking standards could be conducted as follows:

- A sample of teachers in the subject of interest would be asked to submit samples of student work -- tests, examinations, reports, essays, and so forth -- from the previous school year. (This could be done efficiently as part of the surveys of teachers proposed in Chapter 4.) The samples would be of varying quality in the judgement of the teachers, ranging from failing to barely passable to competent to good to excellent.

- A random sample of 50 teachers of a course would be drawn from across the province. They would be presented with the work of 20 fictional students, and asked to assign end-of-term marks to the students.

- The marks assigned would be examined statistically for evidence of among-teacher (among-school) differences in marking standards.

An issue that needs to be addressed here is motivation: Why should teachers agree to participate in a time consuming project of this kind? We have no answer to this question but suggest that two motives might be invoked:

1. professional development; (e.g., The teacher could see by studying the student work in the stimulus materials what information other teachers collect to evaluate student achievement; he/she could be promised a full report of the results of the study, including the distribution of marks assigned by the other participants.)

2. financial remuneration; (e.g., Each participating teacher could be offered an honorarium.)

Information about the procedures these teachers employ in assigning marks to students could also be obtained within the framework of a study of marking standards.

- The teachers could be asked to provide the subjective weights that were applied to each piece of work considered in the marking.

- These weights could be compared to ascertain the extent to which there are differences among teachers about which kinds of student work were important indicators of achievement and which were not.

CHAPTER 7:
CONCLUDING COMMENTS

In this report, we have proposed four studies to address four of the effects that province-wide examinations might cause. To review, these are:

1. effects on marking standards;
2. effects on the curriculum;
3. effects on enrolment trends;
4. effects on public perceptions of education in Ontario.

The proposed studies are of:

1. information in Ministry maintained data bases;
2. information from a survey of teachers;
3. information from a poll of public perceptions;
4. information obtained in a controlled marking situation.

The following table relates each kind of examination effect to each study.

Table 7-1: Tabular Description of the Crossing of Examination Effects and Proposed Studies

Type of Study	Examination Effect			
	Marking Standards	Curriculum	Enrolment Trends	Perceptions
Ministry Data Bases	X		X	
Surveys of Teachers	X	X		
Polls of the Public				X
Controlled Studies of Marking Standards	X			

As noted earlier, several of these studies are also proposed, in a modified form, in the second part of the report on the effects of an assessment model, should that approach, rather than an examination model, be adopted.

7.1 Comments on Priorities

Whenever more than one activity is proposed, it is inevitable that the question of priorities should arise. It is not our intention to recommend a priority-order of studies. Neither do we recommend a priority-order of effects. What we have tried to do is suggest some of the factors that should be considered.

1. Marking Standards. Recall that three facets of marking standards were considered:

    a. Teacher marks. An analysis of the marks in Ministry files would address such issues as inflation in marks and variation from school to school (teacher to teacher) in marking standards. If examinations were instituted, an analysis of information in Ministry files could also assess the influence of provincial examinations on inflation and mark variation. This study would be relatively inexpensive to conduct because no new data would be collected. Unfortunately, this study would yield results that would be difficult to interpret (see Section 2.1).

    b. Teacher procedures for evaluating students. The study of these procedures could be made only if additional data were collected. These data could be obtained as part of a study of the broader issue of the effect of province-wide examinations on the curriculum. This study would give information that a study of marks alone would not: the extent to which teachers differ in the evidence they collect about student achievement and in the weight they place on evidence of different kinds. This study, if done before province-wide examinations were introduced and again after their introduction, would provide information about the extent to which the exam-inations influenced this variation. A side benefit of this study would be that the data collected could be used in in-service training sessions for teachers.

    c. Controlled study of marking standards. A frustrating aspect of any attempt to study teacher marks, as they appear in Ministry data bases, is that each teacher is marking a different class of students. This study would attempt to establish a fixed basis on which to compare teacher marking standards by the artificial means of controlling the information about student achievement that each teacher in the study would be given. Thus, each teacher would receive the same information, although there were no real students to associate with the information. The results of this study should reveal the extent of variation in the achievement teachers expect for a given mark. Unfortunately, this information could not be collected except in a special study devoted exclusively to this purpose. Moreover, it would be a relatively expensive study to conduct. On the other hand data collected during the study would be useful in in-service training programs for teachers on the topic of student evaluation.

2. **Curriculum.** If a study of the curriculum and of instructional methods for a course were conducted both before and after a province-wide examination had been introduced, it should reveal the extent of present variation among teachers in objectives and methods, the extent of variation with the province-wide exam, and the difference between the two situations. The study would show any serious negative effects of examinations on the curriculum and should provide a context for understanding the results of studies of marking standards. The cost of this information would be high. Still, the study of teacher bases for student evaluation could be undertaken in conjunction with this study.

3. Enrolment trends. Several important trends in student enrolments could be detected in a study of existing Ministry data bases. A study of selected trends could be conducted without the collection of additional data. Such a study would be inexpensive, and could be easily done in conjunction with a study of teacher marks. If the effect of examinations on other trends (e.g., the number of low SES students enrolled in courses with province-wide examinations), were of interest, more data would have to be collected. No attempt has been made here to propose a study in which additional data on enrolment trends would be collected.

4. Perceptions. A poll of the Ontario public could study perceptions of the quality of secondary school graduates and of secondary schooling. Such a study could be conducted relatively inexpensively. A negative feature of studies of perceptions is that they are just that -- studies of perceptions. They provide no information about the reality to which the perceptions pertain.

7.2 Limitations

No consideration of priorities would be complete without a review of the limitations of studies of examination effects. Several limitations should be noted:

1. There is a dearth of empirical information available about the effects of externally imposed examinations of the sort we have assumed will be introduced in Ontario. The proposed studies would amount to explorations of virgin territory, and would be conducted without the benefit of the experience of others.

2. The introduction of OSIS entails the development of new curriculum guidelines. If provincial exams for secondary school courses were introduced in the near future, OSIS would produce effects that would be confounded with any effects that the examination system had.

3. If examinations were to be introduced in the near future, the funding of Catholic high schools would have an unknown influence on teaching and learning conditions during the time that provincial examinations were being introduced.

4. Little time remains for the collection of baseline data, which would be essential if the results of the proposed studies were to be interpreted as change due to examinations.

7.3 Closing Remark

Although the proposed study of examination effects on curriculum is the most expensive of our proposals, it is the key study. Investigations of teacher marks and the introduction of an expensive examination system to "raise educational standards" would be sterile and meaningless without studies that would leave us better informed about what is being taught and what is being expected of students in the classroom. To study only mark variation would be to confuse standards with marks. It should be remembered that standards, as defined by the tasks students should be able to do after having taken a course, are set during teaching. Marks are intended to reflect achievement, but they do so in a relative fashion only (one student compared to another in the same classroom), not in any absolute fashion. Marks are best regarded as points on a rubber ruler, a ruler that the teacher can stretch or relax as the situation dictates to produce a result acceptable to the school's administration, if not also to the students, their parents and society at large.

LIST OF REFERENCES

Brown, C. "Ontario's Grade 13 - Guidelines from the Past". Public lecture, Centennial Series, College of Education, University of Toronto, 23 February 1967.

Brown, C. "Ontario's Grade 13 - Concerns for the Future". Public lecture, Centennial Series, College of Education, University of Toronto. 29 February 1967.

Graham, N. "Designing and Marking English Examinations: A Resource Booklet for Scarborough English Teachers". September 1984. (Unpublished report for the Scarborough Board of Education)

Nagy, P. "An Examination of Differences in High School Graduation Standards". Canadian Journal of Education 9 (1984), pp. 276-297.

Traub, R.E., and McLean, L.D. "A Rosy View -- University Admission Officers' Preferences and Expectations for Provincial Examinations". November, 1984. (Unpublished report for the Ontario Institute for Studies in Education)

Traub, R.; Wolfe, R.; Wolfe, C.; Evans, P.; and Russell, H. Secondary-Postsecondary Interface Project II: Nature of Students. Toronto: The Ministry of Education and the Ministry of Colleges and Universities, 1976.

PART TWO

Province-Wide Testing:

Strategies for Evaluating the Impact of Assessments

Part Two

Table of Contents: The Impact of Assessments

1. INTRODUCTION 48
    1.1. Background 48
    1.2. Rationale for an Assessment Program 48
        1.2.1. Advantages 49
        1.2.2. Disadvantages 49
    1.3. Overview 50
    1.4. Facets of an Assessment Program 50

2. POSSIBLE OBJECTIVES OF ASSESSMENT 52
    2.1. Introduction 52
    2.2. Providing Data for Program Confirmation or Revision 52
        2.2.1. Individual information 53
        2.2.2. Teacher involvement and acceptance 53
        2.2.3. Level of difficulty 54
        2.2.4. Process versus product 54
        2.2.5. The time commitment 54
    2.3. Providing Information to the Public 55
    2.4. Reporting Agreement about Goals and Standards 56
    2.5. Promoting Good Testing Practices 57
        2.5.1. Objective-based testing 58
        2.5.2. Higher level objectives 58
        2.5.3. Forestalling misdirected assessments 59
    2.6. Providing Longitudinal and Cross-Sectional Data 60
    2.7. Gaining Teacher Acceptance 61
    2.8. Summary 63

3. THE FACETS 64
    3.1. Involvement of Teacher Federations 64
    3.2. Secure or Open Instrument Pool 65
    3.3. Ancillary Information 65
    3.4. Nature of Student Sampling and Reporting 66
        3.4.1. Student motivation 66
        3.4.2. Use of results 67
    3.5. Time Frame and Grade Levels 67
    3.6. Scope of Data Collection 68
    3.7. The Information Dissemination Process 69

4. THE RECOMMENDED PROPOSALS 70
  4.1. Studies of Marking Standards 70
  4.2. Surveys of Teachers 71
  4.3. Study of Perceptions 72
  4.4. Study of Effects on Policies for Personnel 72
  4.5. Concluding Comments 73

LIST OF REFERENCES 74

CHAPTER 1:
INTRODUCTION

1.1 Background

On 29 June, 1984, the Ontario Ministry of Education sent out a call for proposals to evaluate the impact of province-wide testing. This report is one of three prepared in response to the call.

The term testing suggested two quite different models:

1. the examination model, and
2. the assessment model.

By the examination model is meant a system whereby all students taking a particular course, typically at the high school level and typically at the end of high school, write a common, provincially set examination; a student's mark on that examination counts as a part of his/her final grade in the subject. By the assessment model is meant a system whereby the students in a course write one of several tests; different students in the same class write different tests. The results of an assessment provide information on group (e.g., province or board) levels of achievement, and not on the achievement of each individual student. It is usual for only a sample of the students in a course to be tested in implementations of the assessment model.

In the judgment of Ministry officials and the research team, the two models of testing were different enough that more than one report was appropriate. Thus, the project report is in three parts:

- Part 1 - the potential impact of a provincial examination system

- Part 2 (this part) - the potential impact of a provincial assessment system

- Part 3 - the literature on both models of testing.

Further information about the call for proposals, and a brief historical background, is provided in the introductory chapter of Part 1.

1.2 Rationale for an Assessment Program

The rationale for an assessment program is the need for systematic, group level information on educational achievement. An assessment program can be expected to have effects on many aspects of the educational system and those who participate in it. This chapter begins with a brief overview of advantages and disadvantages of assessment programs.

1.2.1 Advantages

An assessment program, as it is conceived here, with samples of students writing different tests, can cover a broader range of educational achievements than can an examination. One characteristic of the examination model is that only a small sample of the material in a curriculum can be assessed. Since students are able to write for only a limited amount of time, say two or three hours, they cannot possibly be examined on all the objectives of the curriculum. A likely consequence of exams, then, is that they exert a narrowing influence on the curriculum; they make achievement of those objectives which are tested more important than achievement of those objectives which are not. In an assessment program, a sampling procedure permits testing as many objectives as desired, but only on samples of students. In this way, group level data is obtained on achievement of the curriculum, but no individual student is required to write an inordinately long test. Depending on the number of objectives to be covered, the accuracy of results that is desired, and the size of the jurisdiction in question, it is possible that only some of the students in a course will be asked to participate in an assessment.

In an assessment program, it is possible and feasible to administer a variety of non-traditional tests to small samples of students; these tests would yield information on aspects of the curriculum that it would not be feasible to test in an examination. Examples include measures of laboratory skills in science, oral skills in first and second language, and performance skills in music and drama. Due to numbers of examinees and costs, an examination system can include only traditional paper and pencil tests. Note, however, that non-traditional formats, such as the dictation section of the former Grade 13 French exam, are sometimes possible, even in the examination model.

1.2.2 Disadvantages

The major weakness of an assessment program, compared to an examination system, is that it does not yield information by which the achievement level of an individual student can be compared with the achievement levels of all the other students who are tested. Although some assessment programs attempt to provide information for comparing individuals, the methods employed are open to question. [Procedures have been proposed for equating different sets of items, so that the scores attained by the students who took one set of items can be put on the same scale as the scores of the students who took a different set of items (Lord, 1980; Wright & Stone, 1979). We reject this option for two reasons. First, the method is open to question on technical grounds for tests of educational achievement (Goldstein, 1983; Traub, 1983; Traub and Wolfe, 1982). Second, it is difficult to imagine that the public, as well as students and teachers, would readily accept the results of such a test-equating method, were they informed of its characteristics. This is like comparing one athlete's performance in the long jump with another athlete's performance in the high jump for the purpose of saying which is the better athlete. If, as is usually the case, an assessment model with matrix sampling is chosen, that choice should have been made accepting the fact that comparisons of individual students will not, in general, be possible.]

1.3 Overview

The purpose of this report is to outline some of the effects a provincial assessment program could have on the educational system of Ontario, and to recommend procedures for monitoring these effects. As mentioned earlier, this report has been prepared as the companion to a report on the impact of an examination system. Some of the consequences of educational testing might follow from the introduction of either an examination system or an assessment program; thus, some of the monitoring systems we recommend would be useful in either case. For brevity, the details of a monitoring system are presented in only one report. The reader will be referred to Part 1 at appropriate places in this part of the report.

Assessment systems appear in a great variety of forms, but it is not our task to choose among them. The task of designing a system to track the effects of an assessment program is difficult without information on how the assessment program will be implemented. The approach we have taken is to outline in Chapter 2 the major objectives of an assessment program; we also comment on the potential effects of each objective as these are reported in the literature. In Chapter 3, we outline several options that will need to be considered in designing each facet of an assessment program, and we suggest how different choices might affect attainment of the objectives considered in Chapter 2. In Chapter 4, we recommend studies for monitoring possible effects of a provincial assessment program.

In order to set the context for the reader, we briefly preview the facets of an assessment program that are open to choice.

1.4 Facets of an Assessment Program

The effect of an assessment program depends largely on its form. Its designers must make choices about a number of its facets:

1. nature of teacher involvement -- whether and at what stages (e.g., development, interpretation) those who must accept and use the results are involved;

2. nature of the item pool -- whether open or closed;

3. nature of the other information collected to guide the interpretation of achievement levels -- e.g., opportunity-to-learn, teaching strategies, student and teacher demographic data;

4. nature of student sampling and reporting -- whether results will be reported at the classroom, school, board, or provincial level;

5. nature of the curriculum sampling and reporting -- whether achievement will be reported at the item, objective, or domain level; how these "domains" will be defined and

organized; whether and to what extent the assessment will include non-paper-and-pencil methods and will test higher level outcomes;

6. time frame and grade levels covered -- how the assessment will fit into a system of curriculum renewal;

7. scope of the data collection -- whether data will be collected in core subjects only or in other school subjects as well; whether data will be obtained on cognitive achievement only or on social and emotional development as well;

8. nature of information dissemination -- whether reporting will be of measured student performance only, or will include judgment of the value of a given level of achievement.

CHAPTER 2:
POSSIBLE OBJECTIVES OF ASSESSMENT

2.1 Introduction

The design of an assessment program should be guided by its objectives. This chapter outlines the main objectives one might have for a large scale assessment program; it includes commentary on the suitability of each for Ontario and the effects that might follow from attempting to achieve each objective.

The objectives considered are these:

1. to provide data on the strengths and weaknesses of the provincial educational program and the extent of curriculum implementation; this data would be used to confirm or revise policies and practices;

2. to inform the public about student achievement;

3. to provide information about the extent of agreement among teachers concerning goals and standards of education;

4. to provide information about the extent of agreement among teachers concerning teaching strategies;

5. to influence classroom testing practices, and to forestall misdirected and poorly conceived assessment programs;

6. to provide data for longitudinal and cross-sectional comparisons of educational development.

Most important is the objective of gaining teacher acceptance of the assessment program.

The foregoing objectives were derived from the literature on assessment programs in the United States, Britain, Australia, and other parts of Canada, and from discussions among the project team and other educators.

2.2 Providing Data for Program Confirmation or Revision

In a discussion of the United States National Assessment of Education Progress (NAEP), Greenbaum, Garet and Solomon (1977) stated that one of the major objectives of this program is "to obtain meaningful, national data on the strengths and weaknesses of American education [by locating deficiencies and inequalities in particular subject areas and particular sub-groups of the population]" (p. 168). If this objective, translated to the Ontario context, is judged important by the designers of an assessment program, then the Ontario assessment must have several characteristics:

1. Reports of achievement levels must be explainable in terms of the curriculum and instruction -- that is, it must be possible to draw instructional consequences from the data;

2. Coverage of the curriculum must be comprehensive -- that is, assessment must not be limited to the types of objectives that are most easily assessed;

3. A matrix sampling scheme, capable of giving broad coverage, must be used;

4. The schedule of assessment and reporting must be such that:

    a. time is allowed for assimilating and using the information;

    b. a cyclical procedure is followed to allow judgment of the effectiveness of the assessment program in bringing about curriculum change;

5. Those directly responsible for changes in the curriculum (i.e., classroom teachers), must accept the validity of both the results and the instructional consequences that appear to flow from them;

6. Those responsible for either directing or implementing changes in the curriculum must have the skills needed to interpret and evaluate assessment results, and make changes.

Several program characteristics of this objective bear further comment.

2.2.1 Individual information

The results of an assessment that uses a matrix sampling scheme cannot be used to compare the achievement levels of all students. A slight modification in assessment design is possible, so that _all_ students respond to a "core" set of items, as well as to a sampled subset of a larger collection. This modification would make student comparisons possible, but would probably also be perceived as an external examination program. The impacts of examinations are discussed in another part of this report.

2.2.2 Teacher involvement and acceptance

Meaningful change will come about only through collective effort. We comment further on the matter of gaining teacher acceptance later in this report. It is sufficient to note that if assessment results are to be used to guide curriculum change, then close attention must be paid to gaining teacher acceptance of the assessment. Further, if instructional conclusions are to be drawn from patterns of strengths and weaknesses within a curriculum area, there must be understanding of and agreement on educational objectives and on the means for attaining those objectives. This is no easy matter, as the history of attempts to modify curriculum through assessment demonstrates (Gipps & Goldstein, 1983; Greenbaum, 1977).

2.2.3 Level of difficulty

The purpose of the assessment, whether it is to test all curriculum objectives or only a few, and whether it is to prescribe and promulgate standards of achievement or only identify attained levels of achievement, can have an effect on the difficulty of the items used in the assessment. In any curriculum, there will be objectives that virtually all children should achieve, others that a substantial majority should achieve, and still others that only a minority could be expected to achieve. To ignore any of these objectives would be to pitch the assessment program to certain levels of student ability. For example, a program focused on minimum competencies, as was the Australian Studies in Student Performance (ASSP), would say nothing about quality of program for average and above average students. Similarly, a program that takes a norm-referenced approach to instrumentation, such as the British Assessment of Performance Unit (APU), would tend to include items matched in difficulty to the ability of the average student (Power & Wood, 1984).

2.2.4 Process versus product

There are difficulties in trying to draw lessons about instructional process from data about the product of instruction. There can be a great many reasons why students fail to achieve particular objectives:

1. The objectives may be inappropriate for the age group.

2. The objectives may be inappropriate for students who lack enabling skills.

3. The objectives may not have been taught (no "opportunity to learn").

4. The teaching strategies may be inappropriate.

In the absence of other data on process, only the most limited conclusions on curriculum revision can be drawn from assessment data. Thus, a commitment to this objective implies also a commitment to the collection of process as well as product data.

2.2.5 The time commitment

Attempting curriculum reform implies a long-term commitment to an assessment program. Time must be allowed to design the program, gain general acceptance of it, collect baseline data, interpret results, determine instructional consequences, disseminate curriculum modifications, and then go through another cycle of the entire process.

2.3 Providing Information to the Public

Although the goal of providing information underlies almost all of our discussion, the issue of public understanding and use of results deserves separate consideration. The main questions are these: Who constitutes the public? And, what contitutes the information? As candidates for the public, we have:

- students and their parents;
- prospective employers;
- university faculty and admissions officers;
- teachers and school administrators; and
- legislators.

As candidates for information, we have:

- relative achievement of items which test basic skills;
- relative achievement of items which test the higher-order cognitive goals of education;
- relative achievement of items which test the non-cognitive goals of education;
- relative achievement of different schools, regions of the province or segments of the population;
- relative marking standards or expectations set by different schools or teachers;
- relative emphasis placed by different schools or teachers on each curriculum objective;
- relative use made of different instructional strategies and teaching resources.

Having provided these lists, we offer also the following observations:

- The inappropriate use of assessment data by the uninformed for the "evaluation" of teachers and schools is a problem that must be considered; it is a major concern of teachers.
- The public, however defined, is interested in extent of achievement of basic knowledge and skills, as well as the relative preparation of students -- this suggests reporting both on an objective- or criterion-referenced basis and on a norm-referenced basis.

- Universities are interested in prerequisite skills, uniformity of preparation and uniformity of marking -- this suggests an emphasis on the university-bound students, and on fairly tight central control of the curriculum.

- Ministry authorities and legislators will have their interests in program monitoring and policy formation best served by information on (i) achievement of specific objectives, (ii) the extent of variation across schools, and (iii) evidence as to the stability of achievement over time.

- Collection of information on instructional process, achievement of non-cognitive goals and effects of socio-economic context will be very expensive; but it is essential if the assessment program is to avoid the charge that it is concerned with monitoring at the expense of exploration.

- Public education on the use of assessment information must be seen as a necessary part of the information dissemination process.

2.4 Reporting Agreement about Goals and Standards

This objective concerns the interpretation of results when different teachers teach different programs, using different methods and holding different expectations (i.e., setting different "standards"). The interpretation of achievement results is difficult, if not impossible, without knowing teacher goals and practices.

The description and clarification of goals and standards is also important for instrument development. As Greenbaum (1977, p. 162) put it, with respect to the NAEP, it is important "to develop lists of educational objectives that would fairly reflect the aims of American education and serve as guides for the exercise writers". One might argue that because of the discussion it fosters about the goals of education, there is as much to be gained from the development of an assessment program as from its administration. In other words, simply achieving an agreement among educators on what it is important to assess is as important as finding out whether or not the goals have been attained. In Ontario, this difficult task has already been undertaken, to a considerable extent, by the Ontario Assessment Instrument Pool (OAIP) project.

Returning to the point about interpreting achievement information: Even with a common curriculum and common texts, teachers vary a great deal in their judgment of the importance of objectives, in their allocation of time to objectives, and in their strategies for teaching the objectives. In order to interpret the results of an assessment program with a view to continuing good programs and improving weak ones, we must obtain information about opportunity to learn (OTL). This requires data from teachers on the importance of individual objectives, on the time spent teaching them, and on the teaching strategies used. Also, data must be obtained on other factors, such as access to various learning resources.

The issue of different "standards" can be addressed by discussions among teachers of the objectives to be pursued, their operationalization as test items, and the level of accomplishment to be expected of students. It is considerably easier to measure and report achievement, however, than to judge its significance or value. A major problem in any attempt to clarify the standards issue is the policy of age-promotion in the elementary system. This philosophy ensures that most classes consist of students whose common characteristic is that they are about the same age. Such groups contain students on remedial withdrawal, enrichment withdrawal, and the entire range in between. NAEP solved this problem in its first decade of existance by testing groups defined by age rather than by grade. But this practice too has its critics (Power & Wood, 1984), and the NAEP has recently moved to grade-level testing.

2.5 Promoting Good Testing Practices

Another possible objective of an assessment program is the promotion of good testing practices. There are at least three aspects to this objective:

1. encouraging the comparison of an individual's achievement with well-defined goals (criterion-referenced or objectives-based testing) rather than the comparison of one individual's achievement with that of another;

2. encouraging the testing of educational objectives at a higher level than cognitive recall;

3. discouraging inadequately conceived large-scale achievement testing programs.

The first two of these aspects pertain to classroom level testing, while the third concerns school board or university-entrance level.

To influence testing practices at the classroom level, an assessment program must satisfy two requirements:

1. It must be viewed by teachers as a model of good assessment.

2. Teachers must develop the skills needed to draw instructional conclusions from achievement data, whether from a provincial or a classroom assessment program.

These requirements mean that the assessment program must be accompanied by an in-service program, in which correct uses of assessment results are demonstrated.

2.5.1 Objective-based testing

To deal with the public perception of inadequate quality control in the school system, two quite different approaches are possible. One is a norm-referenced approach, in which the performances of individual students or groups of students (classes, schools, boards) are compared to each other. This approach shows how far individuals or groups are above or below the average.

At least three difficulties arise in interpreting norm-referenced information. One difficulty is that those schools above the average are often naively perceived as doing a good job and those below, a bad job. No matter how much a school system or an individual improves, there will still be an average, and some systems or individuals will be above that average, while others will be below. Keeping the concepts of above and below separate from good and bad is not easy. A second difficulty is that norm-referenced data is of little value in improving instruction. Implications for instruction do not follow readily from relative ranking. The third difficulty is that a test cannot always be made to serve both the purposes of norm- and objective-based measurement equally well.

In contrast to a norm-referenced instrument, a well designed criterion-referenced instrument offers a basis for comparing the achievements of students against expectations rooted in the nature of the tasks they are asked to perform. A rationale for this kind of test referencing is that it can possibly yield information to guide the improvement of educational programs. It is never easy to draw instructional and remedial conclusions from the data provided by a criterion-referenced test; but with the aid of supplementary information (e.g., on opportunity to learn), one can at least expect to relate results to goals. At the same time, it must be remembered that goals can be defined in many different ways, and operationalized in items in many more ways. Accomplishment depends on task difficulty, which varies with goal definition and operationalization. It is apparent that care must be taken in drawing instructional implications from achievement data.

The argument that an assessment program can promote objective- or criterion-based testing at the classroom level rests on the assumption that teachers want information about the achievements of their students from which to draw instructional implications. The hope is that if the results of a provincial assessment program are interpreted in terms of tasks and task performance of expectations, then this same approach will be adopted by those teachers who do not already follow it.

2.5.2 Higher level outcomes

The concern has been expressed that current classroom testing concentrates on cognitive objectives at the lower end of Bloom's taxonomy (1956), with emphasis on memorization at the expense of higher level understanding and application. An assessment can be conducted so as to encourage the testing of higher-order objectives. This raises the issue of costs of the three categories of objectives and instruments that can be defined for an assessment:

- The first category includes objectives that can be tested by multiple-choice and short-answer items. These items are the cheapest to score, but the types of objectives they can test are limited to recognition, recall, and perhaps some aspects of higher-order objectives.

- A second category of objectives are those that can also be assessed by paper-and-pencil tests, but which require longer, constructed answers; such answers can only be scored by a teacher or another expert who exercises subjective judgment. These items are more costly to score, but can cover higher-order objectives (e.g., ability to organize material and write it down in a coherent, intelligible way, ability to solve extended problems) that are inaccessible to multiple-choice and short-answer methods.

- The third category, and by far the most costly to assess, consists of objectives that cannot be assessed by written instruments. Examples are performance ability in art and music, oral ability in French and English, laboratory skills in science, and social development. A broadly based assessment program would include instruments for this third category of objectives.

The assessment devices that are expensive to administer and score can be included if they are administered to small samples of students. This probably means that reliable information could be obtained and reported only for large jurisdictions (e.g., the province as a whole).

2.5.3 Forestalling misdirected assessments

To accept this as a goal of an Ontario assessment program, one must also accept the following:

- In the absence of a Ministry initiative, other groups are likely to act.

- The practices described in the following paragraphs are indeed "misdirected."

One category of large-scale assessment that we view as misdirected includes those intended to screen students for admission to post-secondary institutions. Much of current debate over quality control and standards concerns the university-bound student. If entrance exams were to be imposed by the Ontario universities, they could be expected to have a narrowing effect on the curriculum, even for the large number of students who do not go to university. Also, the results for schools of such examinations might easily become, in the eyes of the public, a measure of school quality. This would be a distorted measure because an entrance exam, which is designed to differentiate among an elite group of students, would consist of a preponderance of difficult items; thus, this would be a measure of "school quality for top students".

It is not necessarily true, of course, that the introduction of a provincial assessment program will prevent the universities from designing and administering entrance exams. Whether or not the universities are deterred will depend, among other things, on whether or not a provincial assessment program produces changes that alleviate university concerns about variation among secondary schools in course content and marking standards.

A second category of misdirected assessments includes the attempts by various American states to impose standards on schools. Many of these attempts have taken the form of minimum competency testing programs. Such programs are damaging to the extent that minimum standards become target standards.

A third category consists of the assessments mounted by individual school boards. It is within the jurisdiction of school boards in Ontario to impose system-wide examinations or system-wide administrations of standardized tests. They might do so in response to questions from the public about quality of education. A misdirected assessment in this category might involve the use of commercially available, norm-referenced tests. Several of the problems associated with a norm-referenced assessment program were outlined earlier. Briefly, these include an emphasis on comparisons among the performances of individual students and schools, and a shifting of emphasis in instruction away from the entire curriculum to a portion of it. It might be argued that the present curriculum is too broad and that some narrowing would be beneficial; however, such a narrowing should be based on a better rationale than the content of a commercially available, norm-referenced test.

Apart from this concern, a general difficulty with school board assessments is that some will be well-funded and others will not. Budgetary constraints are difficult to avoid; the problem with low-budget programs is that they inevitably emphasize objectives that are easily assessed. Because this kind of objective is often seen by the public as "basic", the forces that lead to narrowing of the curriculum are very strong indeed.

2.6 Providing Longitudinal and Cross-Sectional Data

Another possible objective for an assessment program is the provision of data for longitudinal and cross-sectional comparisons. The distinction between these two types of comparisons is important. Longitudinal studies follow the same cohort of students (e.g., those starting kindergarten in 1983) for several years to track gains in achievement. Cross-sectional comparisons test the same grade regularly (e.g., grade six every four years) to track trends in achievement at a grade level over time. Public interest is easily aroused in cross-sectional comparisons over time. (Is it true "things were better in the good old days"?) On the other hand, as the literature shows (Husen, 1979; Goldstein, 1983), those involved in a large scale assessment in other countries have come to appreciate, after the fact, the value of longitudinal comparisons.

There are difficulties associated with the collection of cross-sectional data. These difficulties, which need to be considered in planning an assessment program, centre on the problem of making fair comparisons over time. For example, in 1977 a study conducted in the province of Alberta involved the readministration of a standard test originally administered in 1956 (Alberta Education). This study was to compare educational achievement in 1956 with achievement in 1977. At least two difficulties were encountered by those who attempted this comparison: One is that the language of the curriculum had changed so that the wording of some questions was less fair to the 1977 students than it had been to the 1957 students. The other difficulty was that opportunity to learn (OTL) had changed. In twenty years, many additions had been made to the curriculum; this reduced the amount of time spent in 1977 on the topics of the 1956 curriculum that were tested. With these differences in OTL, any comparison of achievement was meaningless. These problems have also been raised in discussions of the decline in the American Scholastic Aptitude Test from 1960 to 1975 (Wirtz, 1977). There are now more subjects in the curriculum; within subjects, the scope and depth of coverage inevitably changes over time; and, the configuration of courses taken by high-school students almost certainly changes over time.

One of the main requirements of valid comparisons of achievement over time is a stable curriculum. But it is not at all clear that a stable curriculum is either possible or desirable. A strong argument can be made that more-or-less continual curriculum renewal occurs simply as a consequence of year-to-year changes in the personnel (students, teachers and officials) engaged in education and that this change is necessary to meet the needs of a rapidly changing society.

Long-term comparisons of achievement should not be undertaken in the spirit of checking present-day quality against a given standard from some point in the past. Rather, they should be undertaken to achieve better and better approximations to a level of outcome widely accepted as a target for future attainment. That is, after the initial results of an assessment program have been used to change curriculum policy, long-term comparisons may be made to determine whether these policy initiatives have been effective.

2.7 Gaining Teacher Acceptance

Many teachers fear assessments. They fear that assessment results will be misused and that the assessment will narrow the curriculum. These fears, left unanswered could result in lack of teacher cooperation. Experience elsewhere indicates that lack of teacher cooperation is a major problem for assessment programs (Nisbet, 1978; Kogan, 1978). The task of gaining teacher cooperation is probably more difficult for an assessment program than for an examination program: Teacher "cooperation" can be forced when individual students' grades are at stake, as they are in an examination program.

Regarding the fear that results will be used for teacher evaluation: There is a fine line between misguided teacher evaluation and appropriate teacher evaluation. Where this line lies is a subject of considerable debate. There are those who consider any teacher evaluation inappropriate, and others who consider any argument from teachers against evaluation as evidence of self-interest.

In addition to fearing the use of test results in the formal evaluations made by supervisors, teachers also fear the use of test results to make inappropriate informal evaluations. These include the ranking of schools or classes. While most educators readily accept that large differences in achievement should be expected among schools from differing neighbourhoods in a city, this is not generally understood by the public. A teacher in an inner-city school, whose students perform poorly in comparison with those of colleagues in suburban schools, may be doing a better job than the suburban teachers.

Regarding the fear that the evaluation of a program would concentrate on a narrow set of objectives: In general, it is considerably cheaper to assess low-level cognitive outcomes using multiple-choice items than it is to assess complex, high-level outcomes, such as writing ability, laboratory skills in science and social development. Because of the financial constraints that are likely to be imposed on an assessment program, there is a real danger that emphasis will be given to those objectives that are most easily and economically assessed.

A related fear is that the objectives that are easiest to assess become de facto most important. For example, if we return to the inner-city school just described, a teacher may be offering an excellent program to children from deprived backgrounds, focusing upon, say, learning skills and social development. Such a program may fare poorly if assessed in terms of low-level cognitive outcomes, but may be a very suitable program for such children. Such a teacher might fear that she would be compelled to offer a program that fails to meet her children's needs simply to satisfy the demands of an externally imposed assessment. Teachers must be assured that the assessment program will not dictate the curriculum to such an extent that professional judgment is obviated.

To allay teacher fears, the developers of an assessment program are best advised to:

- assess broadly, including high- and low-order cognitive objectives;
- involve teachers in the development of the assessment;
- involve teachers in the interpretation of results;
- expend considerable time, energy and resources communicating the assessment results to the public. (By fostering public understanding of assessment data, those responsible for an assessment may be able to allay teacher fears of public misunderstanding.)

2.8 Summary

The foregoing objectives merit consideration in any effort to design an assessment program. Our view is that an assessment program:

- should involve considerable teacher input in its development;
- should include in-service programs for teachers and administrators on the interpretation of results;
- should promote discussion of educational goals and outcomes, in order to achieve consensus;
- should collect other information, such as opportunity-to-learn, teaching strategies, and socio-economic background;
- should emphasize the comparison of achievements with goals rather than the comparison of relative achievements of different groups and individuals;
- should cover as much of the curriculum as feasible;
- should serve as a model for the development of good classroom assessment practices;
- should avoid, as much as possible, any narrowing of the curriculum.

We will now proceed to a discussion of different facets of an assessment program.

CHAPTER 3:
THE FACETS

Many of the studies suggested to monitor the effects of an examination program apply also to an assessment program. The implicit mechanism by which each program might influence the student instruction and evaluation practices of teachers is the main difference between the two programs that affects how they might be monitored. Intervention in the student evaluation process by external examination can be expected to have a direct influence on the instruction and the student evaluation practices of teachers. (e.g., Teachers can be expected to teach to the examination.) In contrast, an assessment program is likely to affect the instruction and student evaluation practices of teachers only to the extent that there is widespread discussion of assessment results and widespread agreement on the importance of what was assessed and on how it was assessed.

Whether assessment results are reported at the system or school level, system personnel will be involved in interpreting and applying local results. This involvement and sampling scheme employed suggest two reasons why more attention has to be paid to smaller school systems than to larger ones under an assessment model:

1. sampling issues -- Results for the smallest boards may not be accurate unless the students in small boards respond to more items than students in large boards.

2. personnel issues -- Small jurisdictions may not have the staff to make full use of the assessment results for program monitoring.

In this chapter, we consider the consequences for Ontario education that might ensue from particular decisions on how an assessment program is implemented.

3.1 Involvement of Teacher Federations

Teacher federations can either be involved or not in the processes of:

1. designing the assessment;

2. interpreting the assessment results.

Whether the professional federations are involved or not is likely to affect the degree to which teachers cooperate in the conduct of the assessment program. Common sense suggests that teacher cooperation is needed in the data collection stage to ensure that students are highly motivated to perform the assessment instruments. Teacher cooperation is also needed to implement whatever changes in curriculum and program are indicated by the results of the assessment. The principals and superintendents of each local board must also show a commitment to change if the instructional program is to be affected by the results of an assessment.

3.2 Secure or Open Instrument Pool

The exercises used in the assessment can either be:

1. created each time, as for a secure test;
2. chosen from an open pool.

This choice will affect the cost of instrument production and the effect that the assessment program has on the curriculum. If new "tests" are developed for every assessment, ongoing development costs will be incurred. If, on the other hand, a large pool is developed, initial costs will be higher, but ongoing costs lower. Estimates of cost are difficult to make for exercise development because curriculum is always evolving, and the pools must be renewed from time to time. (This issue is considered again in the section on long-term comparisons.) But, for example, use of OAIP/BIMO would mean that the instrument development costs would have been borne by another Ministry program, and would not be incurred in the assessment.

In due course, assessment instruments should have an impact on the curriculum, and on what is taught. One might aim to minimize this effect by basing the instruments on just a few objectives. A comprehensive assessment, which reflects the full range of intended curriculum outcomes, should have maximum effect by encouraging teachers to teach the whole curriculum. A secure pool of assessment instruments would have less effect on the curriculum than an open item pool, at least in the short run. Even with secure assessment instruments, however, a publication similar to the Coles' Notes that were developed for the secure Grade 13 examinations of an earlier era might appear and affect the curriculum.

3.3 Ancillary Information

Information could be collected about:

1. teacher goals and expectations;
2. opportunity-to-learn;
3. teaching strategies;
4. socioeconomic status.

These kinds of information would help in the interpretation of achievement results by indicating whether the teachers were teaching the objective associated with an item, whether their teaching strategies were appropriate, and whether one ought to be satisfied, considering the home environments of the children, with a given level of performance. Without such infor-mation, valid interpretation is difficult.

3.4 Nature of Student Sampling and Reporting

Two main issues arise in connection with sampling and reporting:

1. student motivation;
2. use and abuse of results.

3.4.1 Student motivation

Using a matrix sampling system, it is possible to get reliable information on achievement at the board level, for boards of sufficient size, and also at larger-than-board levels of aggregation. If we do not provide test scores that "count" for individual students, we must ask whether the students will try hard enough for the assessment results to reflect their true levels of achievement. Data from the OAIP field trials in high school physics and chemistry suggest that when questions other than multiple-choice were presented, large numbers of students simply did not respond. The problem of motivation is probably more extreme for students of high school age than for younger students.

If reporting is at the classroom or school level, we can make a reasonable, though not entirely convincing, argument that students can be motivated to perform to the best of their abilities. The essence of the argument is that if classroom or school level results are to be made public, the students will do their best because their school's reputation will depend on their performance. In turn, the school's reputation will affect how the student's own marks are viewed.

On the other hand, if reporting is at the provincial level only, then it is difficult to see why students should try hard. Added incentives would have to be built into the program. One way this could be done is to report the performance of individual students to the school for incorporation into the students' grades for the year. This might involve having all students respond to a common set of questions so that comparisons can be made among all the students of their performance of the common questions. Such a system might cause a negative reaction, however, because in effect it would use a form of common examination.

Another approach to increasing student effort is to use only multiple-choice items. A larger proportion of students could be expected to respond to such items than to extended-response items. However the appearance of greater effort may be illusory, since responses to multiple-choice items can be made at random. As well, the types of objectives covered by the assessment would be limited to those that can be tested by multiple-choice items.

One more alternative is worth considering as a means of increasing motivation. The assessment could be constructed so that in each participating school, all students would write the same set of items on a particular topic. Different sets of schools would be assigned different sets of items. In this scheme, the teacher would be returned marks that could be counted toward the student's grade for the course. School results could be aggregated to give provincial data. Because the scores would count, students would have reason to do their best.

Several difficulties arise from this last proposal. First, a teacher might argue that the assessment instrument chosen for his/her school was not appropriate for the way the particular topic had been taught. Second, schools might have to be informed in advance of the topics on which they would be tested (although it is reasonable for teachers and schools to be expected to prepare for an assessment of the entire curriculum for the course). This would lead to cramming on the topic to be tested and to the neglect, to some extent, of the rest of the curriculum. Thus, achievement results would be inflated; except for the very largest boards, the sampling scheme would provide provincial (or regional) results only, and even then, sampling errors would be relatively large.

3.4.2 Use of results

Unless the results of an assessment are made public, it is difficult to see how the public's view of the quality of education could be changed by the assessment. If, however, the results are made public, unfair comparisons might be made of boards (or schools, if reporting is at that level). Also, public reporting may lead to concentration on a subset of the goals of education, that subset on which a board's students performed relatively well.

If high school results were to be reported, either by the Ministry or the boards, schools could be ranked on the basis of their aggregate results, and the ranks then used to weight the school marks assigned to students. This use would affect a student's chances of entering college or university.

3.5 Time Frame and Grade Levels

Decisions on these aspects of an assessment program will determine how often the assessment of a particular subject for each grade is repeated; hence the decisions will determine how the results can be used in cyclical review and development of curriculum. Experience in other countries (Greenbaum, Garet & Solomon, 1977; Gipps & Goldstein, 1983) demonstrates that much information can be lost by too short a time cycle. Those responsible for the assessment find themselves organizing for the next administration before information from the previous one is fully analysed, understood, and acted upon. As well, the grades chosen for assessment of a subject should be well spaced; the curriculum consequences of an assessment at a given grade almost certainly carry over to adjacent grades. To see how this could happen, suppose mathematics is the subject for assessment in Grades 4, 6 and 8. Besides affecting the curriculum of those grades, the assessment will almost certainly affect the curriculum of the intervening grades. Thus, if the assessment at the Grade 6 level shows a weakness in, for example, fractions, this might lead to curriculum changes in both Grades 5 and 6. It might be unnecessary to assess fractions at the Grade 8 level before the impact of the changes in the curriculum of Grades 5 and 6 had time to appear.

3.6 Scope of Data Collection

There are two inter-related issues here:

1. the types of objectives assessed
2. the curriculum subjects assessed

As noted earlier, there are three combinations of instrument types and objectives that can be used in an assessment:

1. multiple-choice and short answer items -- These are the cheapest to administer, but provide the narrowest coverage of curriculum and cognitive objectives.

2. the foregoing item types plus longer essay formats -- Such devices provide greater coverage but at higher costs for scoring.

3. both the foregoing types of instruments, plus assessment instruments and procedures that do not involve paper-and-pencil, for example, instruments and procedures aimed at oral skills in French, laboratory skills in science, and performance skills in art or music -- These kinds of assessment instruments and procedures provide the most complete coverage of curriculum and cognitive objectives, but at the highest cost.

There are no technical barriers to the conduct of a full assessment, that is, an assessment in which non-paper-and-pencil instruments are included. For example, some years ago, the State of Michigan did an assessment of the elementary music program across the state. Assessors went to sampled schools, tape recorders in hand, and took samples of the children's abilities to repeat rhythm patterns, to sing simple tunes, and other such skills.

The cost of a full assessment, is such that neither school nor board-level data could be obtained for the non-paper-and-pencil instruments. For these types of instruments, only provincial data would be feasible. The use of multiple-choice assessment instruments only would lead to narrow coverage. A good compromise between coverage and cost would involve (i) multiple-choice testing of most students with reporting at the school level, (ii) further written devices administered to a few students in each school, with reporting at the board level and (iii) instruments of other types administered to a very small sample of students/schools, with reporting at the provincial level. It might be possible to build in the option of local participation in the full assessment, but at local expense.

Two competing considerations affect the decision about which subjects should be assessed. One is the danger of creating first-class and second-class subjects. The other is the public's need for information about achievement in "the basics". For example, there may well exist a public demand for information about achievement in mathematics and first language, but there is probably not a public demand for information on achievement in art and music. This suggests that all subjects should be assessed in some fashion on a rotating basis. A broad assessment program will foster public awareness of schooling beyond the basic subjects. It will also ensure that the curriculum is not unnecessarily narrowed by a focus on core subjects.

3.7 The Information Dissemination Process

One view of the reporting of assessment data is that it should be neutral and factual. Another is that it should contain expert views on the value of the results, that is, whether those involved consider the results in a particular area to be "poor", "fair", or "good". Encouraging teachers to discuss the results of an assessment can both heighten their awareness of the assessment and increase their use of the results in the classroom. Encouraging members of the general public to discuss the results is likely to heighten public awareness of the assessment, and of the problems of interpreting school achievement. Any increase in the level of public understanding will be beneficial.

CHAPTER 4:
THE RECOMMENDED PROPOSALS

To review and organize: In the part of the report on examination effects, we proposed six studies:

- marking standards, both without and with corollary information; This study would draw on data in Ministry records. In the first case (without corollary information), it would examine trends over time in variations in teacher-assigned marks, and in the second case (with corollary information), it would examine the relation between teacher-assigned marks and provincial exam marks.
- enrolment trends; Again data contained in Ministry records would be used.
- the bases on which teachers assign marks; This study would use data collected from teachers.
- curriculum effects; Again data collected from teachers would be used.
- the perceptions that various publics have of education; Survey methods would be used.
- teacher marking standards; This would involve a sample of practicing teachers in a controlled study.

Of these, it is doubtful that an assessment would affect enrolment trends, so this study is not considered here. The other proposals are considered in turn. Finally, we propose an additional study of effects on personnel evaluation policies.

4.1 Studies of Marking Standards

In our consideration of examination effects, we described the contents of three Ministry data bases that would be useful in monitoring the effects of a provincial examination system. We then proposed a study of marking standards, both with and without corollary information, and a study of enrolment trends. It is doubtful that enrolment trends would be affected by an assessment program, but it seems possible that an assessment program could affect teacher marking standards. The Ministry data bases are limited to the SSHGD level for information on individual students; they would be of very little use in tracking the effects of an assessment program on marking standards. Thus, the monitoring procedures recommended for an assessment are different from those suggested for examinations.

A question to be answered is whether the existence of an assessment program influences the standards that teachers refer to and apply, either explicitly or implicitly, when they mark examinations, projects, etc. The studies recommended in response to this question are of two types:

- an ongoing study of the marks that are awarded without corollary information -- this is the equivalent of the first proposal from Part 1 of this report;
- a controlled study of teacher marking standards -- the equivalent of the sixth proposal from Part 1 of this report.

The first of these studies would track, over time, the marks awarded by a large sample of teachers to the students in a restricted set of grades (say 6, 8, and 10). The study would cover several basic subjects; mathematics and language arts/English are suggested. Marks would be collected from the sample of teachers every two years, beginning before the introduction of an assessment program and continuing for a period of time thereafter; they would be examined for evidence of inflation/deflation and increasing/decreasing variation over schools.

There are three main differences between this proposal and its equivalent in the part of the report on examinations:

- The data-base for this study would be more limited; fewer teachers and subjects would be examined.
- The data would be harder to get because Ministry data bases would not be available.
- The examinations study would involve only SSHGD level courses whereas the study proposed here would involve several other grades.

Only a study of teacher marks would be conducted; no comparison of teacher marks and exam scores for individual students could be made. The object would be to see if teacher marks vary from school to school and change over time. Use of a relatively small sample would limit costs. Since the data would have to be collected from individual schools -- they could not be obtained from extant Ministry files -- costs would be increased.

The controlled study of teacher marking standards that was proposed in Part 1 of this report is relevant, because an assessment may be expected to affect the way teachers evaluate student achievement. This proposal is suggested here, with no modifications in objectives or design.

4.2 Surveys of Teachers

Two studies were proposed in our consideration of the effects of examinations that would require the collection of data from teachers:

1. a study of the basis for teacher marks -- the types of instruments and procedures used for collecting evidence of student achievement, and the procedures used in combining the evidence;

2. a study of the influence of the examination program on the curriculum, as taught in the classroom.

An objective of an assessment program might be to influence the basis on which teachers evaluate student achievement. For this reason, the study of the basis for teacher marks that is proposed in Part 1 of this report is offered unchanged for the study of assessment effects.

The effect of an assessment program on the curriculum might be expected to be less than that of an examination program. This effect should nevertheless be monitored. Thus, the study proposed in Chapter 4 of the report on examination effects is suggested again here.

Because the influence of an assessment program depends very much on the follow-up activities, and because these depend on the availability of the resource staff, the effects of an assessment on the basis for teacher marks and on the curriculum might be different for smaller boards than for larger boards.

4.3 Study of Perceptions

One risk of an assessment system is that no one may know it exists. An examination system affects the grades of every student in the examination course and must be acknowledged and dealt with by the school system and public; such is not the case for an assessment program, which reports achievement only at aggregated levels. Chapter 5 of the report on examination effects proposed a study of public perceptions, including the perceptions of teachers. That proposal is equally applicable for studying the effects of provincial assessments on public perceptions, including teachers' perceptions.

4.4 Study of Effects on Policies for Personnel

Assessment results might affect policy on personnel evaluation and professional development. To study both effects a sample of boards and of schools within boards would be required; however, the data collected would be anecdotal, generally small in volume, and not amenable to quantitative analysis.

We suggest two major data sources:

1. interviews, probably by telephone, of a sample of officers of the teachers' federation and of officials of school boards, concerning the uses being made of assessment results;

2. systematic perusal, using the provincial government clipping service, of all Ontario daily and weekly newspapers, for reports of local use of assessment results.

4.5 Concluding Comments

This report has outlined the possible objectives of a provincial assessment program, and the aspects of such a program that are open to choice by the program designers. Although the match between possible objectives and choices within the various aspects is not clearcut, we have attempted to discuss the effects that might follow from the various choices.

At the conclusion of Part 1 of this report, on examination impact, we commented on factors that should be considered in setting priorities among the studies that were recommended for monitoring the effects of an examination program. These comments hold as well for studies to monitor the effects of an assessment program. And again, we wish to stress that effects on the curriculum are most important. They should be monitored most closely.

LIST OF REFERENCES

Alberta Education. Edmonton Grade 3 Achievement Study: 1956-1977 Comparisons. Edmonton: Alberta Education, 1977.

Bloom, B. Taxonomy of Educational Objectives, Handbook I: Cognitive Domain. New York: McKay, 1956.

Gipps, C., and Goldstein, H. Monitoring Children: An Evaluation of the Assessment of Performance Unit. London: Heinemann Books, 1983.

Goldstein, H. "Measuring Changes in Educational Attainment Over Time: Problems and Possibilities". Journal of Educational Measurement 20 (1983), pp. 369-378.

Greenbaum, W.; Gartet, M.S.; and Solomon, E.R. Measuring Educational Progress: A Study of the National Assessment. New York: McGraw-Hill, 1977.

Husen, T. "An International Research Venture in Retrospect: The IEA Surveys". Comparative Education Review 23 (1979), pp. 371-385.

Kogan, M. "The Impact and Policy Implications of Monitoring Procedures". In Accountability in Education edited by T. Becher and S. Maclure, pp. 113-126. Windsor, U.K.: NFER Publishing, 1978.

Lord, F.M. Applications of Item Response Theory to Practical Testing Problems. Hillsdale, N.J.: Lawrence Erlbaum, 1980.

Nisbet, J. "Procedures for Assessment". Accountability in Education edited by T. Becher and S. Maclure, pp. 95-112. Windsor, U.K.: NFER Publishing, 1978.

Power, C., and Wood, R. "National Assessment: A Review of Programs in Australia, the United Kingdom, and the United States". Comparative Education Review 28 (1984), pp. 355-377.

Traub, R.E. "A Priori Considerations in Choosing an Item Response Model". Applications of Item Response Theory edited by R.K. Hambleton, pp. 57-70. Vancouver, B.C.: Educational Research Institute of B.C., 1983.

Traub, R.E., and Wolfe, R.G. "Latent Trait Theories and the Assessment of Educational Achievement". In Review of Research in Education edited by D.C. Berliner, pp. 377-435. Washington, D.C.: American Education Research Association, 1981.

Wirtz, W. On Further Examination. New York: College Entrance Examination Board, 1977.

Wright, B.D. and Stone, M.H. Best Test Design. Chicago: MESA, 1979.

PART THREE

The Impact of Testing Student Achievement:

A Review of Literature on the Impact of Testing

Part Three

Table of Contents: The Review of Literature

A PREFACE TO THE REVIEW OF LITERATURE 78

SECTION I: THE IMPACT OF EXAMINATIONS 80

1. THE IMPACT OF EXAMINATIONS: AN INTRODUCTION 80

2. LIMITS TO GENERALIZABILITY 82
   2.1. Confounding Variables 82
   2.2. An Advance Organizer for Discussion of Effects 83

3. POTENTIAL EFFECTS ON INDIVIDUALS 84
   3.1. Effects on Students 84
   3.2. Effects on Teachers 87
   3.3. Effects on Other Public Groups 89

4. POTENTIAL EFFECTS ON THE TEACHING-LEARNING PROCESS 91
   4.1. Effects on Classroom Interactions and Activities 91
   4.2. Effects on the Implemented Curriculum 92

5. POTENTIAL EFFECTS ON INSTITUTIONAL POLICIES 95

6. CONCLUSIONS 98

LIST OF REFERENCES 98

SECTION II: THE IMPACT OF ASSESSMENTS 104

1. THE IMPACT OF ASSESSMENTS: AN INTRODUCTION 104
   1.1. An Overview of Implemented Assessment Programs 104
      1.1.1. International assessment 105
      1.1.2. National assessment 105
      1.1.3. Intranational assessment 106
   1.2. An Advance Organizer 107

2. IMPLICATIONS OF PLANNING DECISIONS 108
   2.1. Purposes of Assessment 108
   2.2. Contextual Measures 110
   2.3. Technical Issues of Assessment 112

3. POTENTIAL EFFECTS ON INDIVIDUALS 118
    3.1. Effects on Students 118
    3.2. Effects on Teachers 119
    3.3. Effects on Other Public Groups 121

4. POTENTIAL EFFECTS ON THE TEACHING-LEARNING PROCESS 124
    4.1. Effects on Teaching 124
    4.2. Effects on the Implemented Curriculum 127
    4.3. Effects on Evaluation of the Teaching-Learning Process 127

5. POTENTIAL EFFECTS ON INSTITUTIONAL POLICIES 130

6. CONCLUSIONS 133

LIST OF REFERENCES 134

APPENDICES 136

Appendix A: A Selectively Annotated Bibliography on the Impact of Examinations 136

Appendix B: A Selectively Annotated Bibliography on the Impact of Assessments 163

A PREFACE TO THE REVIEW OF LITERATURE

The following review is part of a larger project funded by the Ontario Ministry of Education, the objectives of which were:

- to "review the literature";
- to develop "desirable and feasible options for a system or systems of monitoring [the effects of procedures] for evaluating...student achievement in Ontario schools." (Request for Proposals for Research, p. 3.)

The evaluation of student achievement has been considered from two testing perspectives, the examination model and the assessment model. Both models were identified as being feasible in the Ontario school system. It could not be assumed that the potential impact on the system would be the same for each model; thus, the impacts of examinations and assessments have been considered separately. The final project report consists of three parts. These are:

Part One, Strategies for Evaluating the Impact of Province-Wide Examinations,

Part Two, Strategies for Evaluating the Impact of Province-Wide Assessment, and

Part Three, The Impact of Testing Student Achievement: A Review of Literature on Examinations and Assessments.

This document is Part Three, the review of literature on the impact of testing student achievement. This review was conducted with three objectives in mind:

- to identify programs in which effects of testing have been monitored;
- to identify types of effects which occur as a result of testing;
- to assess the applicability of the findings reported in the literature to education in Ontario.

The review of literature was conducted in three phases:

1. A search of several major educational data bases and a request for suggestions from several educational researchers in the English-speaking world led to the compilation of a bibliography of journal articles, books, unpublished reports, newspaper articles, and ERIC documents. Most of these were released during the past ten years.

2. Each document turned up in the search was reviewed, and those judged relevant were annotated.

3. A summary and synthesis of the information derived from the literature search is contained in this document.

Section I of this review outlines the potential effects of examinations, while Section II outlines the potential effects of assessments. In each section these are discussed under the following headings:

1 effects on individuals;

2. effects on the teaching-learning process;

3. effects on institutional policies.

Part Three contains two appendices, one for each testing model. These appendices consist of annotated bibliographies of journal articles, books, unpublished reports, and other related materials.

To summarize the review: We failed to find reports of testing situations that parallel Ontario's. The analyses described in the literature were not designed and conducted as an integral part of a testing program, but were undertaken after the fact, when problems were seen to have arisen. It seems that an impact monitoring system that is integral to a testing program either has not been attempted before or is unreported in the literature available to us. The development and introduction of an impact monitoring system for an Ontario province-wide testing program would be an important innovation in testing practice.

SECTION I
The Impact of Examinations

CHAPTER 1:
INTRODUCTION

"Evaluation is a two-edged sword which can enhance student learning and personality development or be destructive of student learning and personality development. It can have positive or negative effects on teachers, curriculums, and school systems .... it is possible to use evaluation procedures wisely, so that they may have a beneficial effect on learning and teaching. This is a matter of designing and using evaluation with a clear awareness of its possible effects and with a sensitivity to the ways in which the evaluation will be perceived by students, teachers, school authorities, and school patrons or the public" (Benjamin Bloom, 1969, p. 45).

This discussion of examination effects assumes the following about an "examination":

- An examination is an instrument designed to measure learning. The exam might be intended to reflect the degree to which a student has learned the content -- the knowledge, skills, and understandings -- of the curriculum of an academic course, or the degree to which a student has mastered competencies in the content a program of courses.

- An examination for an academic course or program is usually administered to every student enrolled in the course or program at the same time and after the curriculum has been taught.

- An examination is usually identical for every student enrolled in the course/program; this enables a comparison of the performance of students in the course/program.

- An examination may be set by an individual teacher to examine one class of students, or it may be set by an examining committee at a provincial, state or national level for the purpose of examining every student in the province, state or nation.

- An examination score may be used as a criterion in decisions about certification and selection for academic programs of study.

1.1 The Literature

1.1.1 Types of examinations

Two types of examinations dominate discussions in the recent literature: minimum competency tests, and grade-specific or form-specific external examinations. United States discussion papers have concentrated on the dangers of minimum competency testing programs, and have raised validity, ethical and legal issues (e.g., McClung, 1978; Lewis, 1979; Popham and Lindheim, 1981;

Airasian and Madaus, 1983). A few empirical studies have dealt with racial (black - white) differences in passing rates (e.g., Trusz and Parks-Trusz, 1981; Serow and Davies, 1982).

Grade-specific or form-specific external examinations have a long tradition in Britain and other Commonwealth countries, including Canada. In Britain and Australia, recent research has been done to examine the utility of exam scores for predicting university grades, and to examine the relationship of exam scores to socioeconomic status (e.g., Glossop and Roberts, 1980; Crum and Parikh, 1983). Much of the recent Canadian literature has come from Alberta, where departmental exams were dropped in 1973, and reintroduced in 1983. Related studies of public impressions and mark variations have been reported by Dumont (1977), Ratsoy (1983), and Reid (1978).

1.1.2 Types of reports

Three categories of reports were found: large-scale empirical surveys, small-scale correlational studies, and discussions. Only the last were found in any number. Of the large-scale surveys, two were U.S. studies. These were designed to assess the instructional effects of tests and other evaluation methods (Barnette and Thompson, 1979) and teachers' attitudes to testing (Herman and Dorr-Bremme, 1983). In a third large-scale survey, the effects of introducing standardized testing in Ireland were studied (Kellaghan, Madaus and Airasian, 1982). In Canada, surveys by Dumont (1977), Reid (1978) and Ratsoy (1983) for Alberta Education dealt with mark variations, public impressions of standards, and the reintroduction of compulsory Grade 12 exams.

Of the small-scale correlational studies, two were conducted to examine whether high school exam marks predicted later performance at university (Crum and Parikh, 1983; Dunn, 1982), and one examined how teachers used test results (Salmon-Cox, 1981).

Most of the citations in the annotated bibliography are discussion papers, most from the United States on the topic of minimum competency testing. These papers list many possible types of effects of testing (e.g., on the curriculum) and offer discussions of measurement issues (e.g., validity of tests). The arguments advanced in most papers are not substantiated by empirical data. Neither do these papers describe a research methodology for assessing the effects that are discussed.

CHAPTER 2:
LIMITS TO GENERALIZABILITY

Airasian, Madaus and Pedulla (1979) and Madaus and McDonagh (1979) considered two characteristics of a testing program that determine much of its impact. These characteristics are (i) degree of outside control over the test and (ii) importance of test results on life chances.

With respect to the first characteristic, it appears that the Ministry of Education would impose a high degree of external control over provincial examinations for graduating students. The effect of the importance of exam results for students, is less clear. This effect would depend on such factors as:

- the extent to which results count toward the final course mark;
- the reporting scheme that is adopted (i.e., whether the test score is reported separately from the teacher's mark);
- the purposes served by the test results.

It is difficult to predict how provincial exam results might eventually be used. The effect of such examinations on the lives of Ontario students, that is, on students' post-secondary opportunities, might depend on an unresolved issue: the way post-secondary institutions use the exam score in admissions procedures.

2.1 Confounding Variables

Two additional considerations are important in assessing the impact of examinations:

1. how the testing program is implemented;
2. whether both the pre- and post-test phases of the program will be considered.

With respect to implementation, some attention has been paid in the U.S. literature to how minimum competency testing programs are introduced. Tyler et al (1978) reported that the implementation of the Florida testing program was faulty in the following respects: "in lack of adequate communication, lack of careful consideration of all important effects of such a program, lack of planning to try to reduce or eliminate undesirable effects, and lack of decentralization to the school building level of decisions that seriously affect teachers, students and parents" (p. 33). Tyler et al went on to say that the testing program may not have been appropriate for a large segment of the first tested graduating class (the Black and poor) who experienced a very high failure rate, and who may have been "sacrificed for the purpose of rapid implementation" (p. 35). Madaus and McDonagh (1979) concluded that other states should learn from Florida's experience and not introduce minimum competency graduation tests abruptly and without taking account of the educational history of the first class to be affected. Popham and Lindheim (1981) discussed the fairness of minimum competency tests, and commented that in the early stages of a minimum competency testing program, teachers will not have had time to change the focus of their instruction to fit the testing emphasis. The advice of these authors would be to introduce a new testing program slowly, so it can be done well.

Bloom (1969) described three phases of a testing program: the pre-examination phase, the examination phase, and the post-examination phase. Different effects might appear in different phases, so that the breadth of the monitoring process should be considered. Among the effects noted for the pre-examination phase are:

a) student and teacher anticipation and anxiety;

b) student and teacher preparation (e.g., teaching to the test);

c) general narrowing of curriculum objectives.

During the second examination phase, there may be important effects on students' self-esteem (e.g., sense of accomplishment, fear of failure). Post-examination effects may be very profound and long-lasting, depending on the uses made of the examination results. Failing the exam or attaining a score that falls short of the criterion for admission to a post-secondary institution are examples of such effects.

2.2 An Advance Organizer for Discussion of Effects

The preceding discussion has provided a framework within which to consider the potential effects of examinations. The rest of the paper consists of four chapters devoted to different aspects of examination impact.

- Chapter 3 outlines some potential effects on individual students, teachers and other public groups, as reported in the literature.

- Chapter 4 is a discussion of potential effects on the teaching-learning process, specifically of effects on classroom interactions and the implemented curriculum.

- Chapter 5, on the potential effects on institutional policy, covers evaluation, post-secondary admissions, and communications policies.

- Chapter 6 is the concluding chapter; it contains a summary of general and specific conclusions about the impact of examinations.

CHAPTER 3:
POTENTIAL EFFECTS ON INDIVIDUALS

> "Perhaps the main point to be made about the effect of examinations is that it is largely a perceptual phenomenon. That is, if students, teachers, or administrators believe that the results of an examination are important, it matters very little whether this is really true or false - the effect is produced by what individuals perceive to be the case" (Bloom, 1969, p. 44).

3.1 Effects on Students

3.1.1 Test anxiety

> "Test anxious people see evaluational situations as difficult, challenging, and threatening and themselves as ineffective in coping with academic challenges. They focus - sometimes obsessively - on the undesirable consequences of personal inadequacy. Their self-deprecating thoughts are strong and interfere with orderly problem solving. Test anxious people frequently expect and anticipate failure and loss of regard by others". (Sarason, 1983, p. 133).

Test anxiety is raised in discussions of exam effects on students. The concern of school personnel (teachers, principals and counsellors) is that the scores of test-anxious students do not reflect those students' true skills level. Test anxiety affects individual students to different extents; while some students experience a mild form of anxiety that puts them at their competitive peak (Ligon, 1983), other students can suffer from a severe form of anxiety that can cause debilitating effects (Sarason, 1983).

Several articles discussed ways of counteracting test anxiety; these might be considered secondary effects of examinations. Ligon (1983) discussed the need for teachers and principals to prepare students over the course of the school year to take standardized tests; this would minimize the influence of factors that prevent best performance on tests, including test anxiety. Ligon (1983) believed that the heart of the problem was the unfamiliarity of the testing situation, and that the solution lay in giving students experience with "the features of standardized tests that are uncharacteristic of regular classroom instruction and teacher-made tests: multiple-choice items, the wide variety of item formats, time limits, separate answer sheets" (p. 20).

Atkinson (1981) discussed the role and activities of school counsellors in protecting students from test anxiety and other potential side-effects of state-imposed, minimum competency testing (MCT) programs. Atkinson considered test-anxious students to be disadvantaged when expected to demonstrate skills on minimum competency tests, particularly when "the provision of sanctions against those who fail is likely to increase their anxiety and lower their performance" (p. 23). He concluded that, as MCT programs become more prevalent, counsellors need to be aware that test anxiety is a condition that may influence student test performance and "be prepared to advocate on behalf of students affected by [this condition]" (p. 26).

Several other factors are mentioned in conjunction with test anxiety. Bloom (1969) commented that the examination situation, in addition to arousing anxiety, can also cause frustration and self-doubt; Ligon (1983) discussed anxiety as one of four factors that prevent best test performance; the other three are carelessness, confusion and poor use of time.

3.1.1.1 The fear of failure

"Examinations are more threat-provoking than most educational settings. The formality and time-pressures conspire with the importance of good results to shift the balance between hope for success and fear of failure firmly towards the latter" (Entwistle,1981).

Fear of failure is one aspect of test anxiety. It can arise in any situation where a person can succeed or fail, or interpret performance in those terms; exams are but one such situation.

The feeling of failure is fear-of-failure realized. It can arise because of a failed examination or as a consequence of not having qualified to take an examination. Among the concerns expressed most frequently by the public in response to the introduction of external examinations in Alberta were the welfare of unsuccessful students and such consequences of examinations as "an increased high school dropout rate, and negative effects on self-concept, rebelliousness and other psychological aspects, particularly as these relate to non-academic students" (Ratsoy, 1983, p.25g). Reporting on the Chinese National Examination System introduced in 1977, Epstein noted that failed candidates developed a sense of failure and many contemplated suicide. Questions then arise about the responsibility of the school to counsel students who fail or do poorly on examinations, and therefore may have to choose different careers (Atkinson, 1981).

A summary of potential effects of test anxiety might include the following:

a) Test anxiety could be one of the more immediate consequences of the introduction of an examination system in Ontario, because students are inexperienced at test-taking (e.g., Bloom, 1969; Atkinson, 1981; Ligon, 1983).

b) To counteract test anxiety, exam-coaching classes and guidance counselling sessions may be introduced in schools (e.g., Atkinson, 1981; Ligon, 1983).

c) Secondary effects due to fear of failure might be an increase in the percentage of students entering the general stream of high school courses and an increase in the percentage of students dropping out of the system (e.g., Ratsoy, 1983).

d) Later post-examination effects, such as loss of self-esteem, may be felt by students who actually fail the exam or fail to meet the requirements to continue towards their desired vocation (e.g., Bloom, 1969; Epstein, 1982).

e) The reality of failure may result in such secondary effects as additional vocational counselling for these groups (e.g., Atkinson, 1981).

3.1.2 Attitudes towards learning

"Many pupils going into an examination will still be rehearsing in (short term memory) an array of facts, formulae, and outline answers...It is not only during the examination, but throughout the revision period, that pupils may feel pushed towards memorization...A common experience of many pupils who have relied on over-learning to pass an examination is that the 'slate' is enthusiastically 'wiped clean' of the knowledge, once the anxiety is removed" (Entwistle, 1981, p. 261).

Student attitudes towards learning may change if a provincial examination is introduced. Epstein (1982) reported that, after the new Chinese examination system was introduced in 1977, students assumed a more passive role and adapted to a preconceived learning environment (p. 86). Bloom (1969) discussed the competitive, "beat the system" attitude necessary to survive in an examination system where "passing the examination by cramming, studying the tricks of examiners, memorizing material just for the examination, and other examination-taking strategies are separable from learning the subject" (p. 46).

In a Scottish investigation, Sharp and Thomson (1984) looked at the effect of external exams on students' study habits. Analysis of examination marks and student responses to an attitude questionnaire led to the finding that the students who earned better scores on the examinations were more positively motivated toward school work; they were less defensive about their teachers, perceived more support or pressure from home, were more active and more independent learners, and were better able to cope with pressures of examinations.

"It may be that the importance of examinations in contemporary society is sufficiently stressful to generate a coping response in which the tone is primarily cognitive and expedient rather than committed" (Sharp and Thomson, 1984, p. 50).

Some potential effects related to student attitudes towards learning are these:

a) Students' attitudes to learning may become more influenced by their need to succeed in the examination, and less influenced by their interest in the particular subject (e.g., Bloom, 1969).

b) The perceived importance of examination success may trigger such secondary effects as improving student motivation to study (e.g., Sharp and Thomson, 1984).

3.1.3 Classifying students

"It is easy to forget that the examination system was originally designed for a relatively small elite of pupils and that it has grown in a somewhat topsy-turvy manner to cover a much larger proportion of pupils than was originally intended" (Gray, 1981, p. 33).

One effect of examinations has been the streaming of students by ability and the consequent "emergence of schools within schools, with pupils grouped according to whether they will likely be certified" (Madaus and Airasian, 1977, p. 84). In the new Chinese examination system, ability grouping was reinstituted at every age level, as pressures increased for compulsory testing at these various levels (Epstein, 1982). Passmore (1983) reported that the British practice of streaming students into exam and non-exam groups led to the exam groups receiving more than their fair share of resources and the non-exam groups displaying greater absenteeism, disruptive behaviour and lack of motivation.

Recent articles in the London Times Education Supplement voiced concern for the neglected students, those who have demonstrated insufficient ability to sit for the exams (e.g., Makins, 1983). The needs of students outside the academic, exam-oriented program was raised as a potential problem by the respondents to a public opinion survey about Alberta Education's proposal to adopt compulsory Grade 12 exams (Ratsoy, 1983). Ratsoy further noted the public concern that the needs of non-academic students be met by establishing either separate examinations or a different type of high school diploma. Passmore (1982) commented on the situation in Wales; here some schools deliberately avoid the traditional goal of teaching to the examination in order to devise educationally successful courses for their less able high school students (p. 10).

Province-wide examinations may have the following effects:

a) There may be more rigorous streaming of high school students, according to ability or chance of passing the external exam (e.g., Madaus and Airasian, 1977).

b) There may also be more rigorous ability streaming in junior high schools, accompanied by increased use of standardized test scores as the basis for streaming (e.g., Epstein, 1982).

c) When priority is given to students in academic programs that end in examinations, non-academic students, their parents and the general public may be concerned that students in non-academic programs deserve equal consideration (e.g., Passmore, 1983; Ratsoy, 1983).

3.2 Effects on Teachers

Most of the recent reports on how teachers deal with testing come from research on the use of minimum competency testing in classrooms. [It is worth noting that minimum competency testing, in which the test items are not closely tied to the curriculum, has been found to have little impact on teachers in the classroom (e.g., Madaus et al, 1979; Kellaghan et al, 1982)]. Despite major differences between this kind of testing and the school leaving examinations that are attracting attention in Canada, lessons may be drawn from the attitudes of teachers toward minimum competency tests.

It appears that teachers trust their own judgment more than a score on a minimum competency test. Yeh (1978) found that teachers believed test scores to be more a product of test taking skills and student motivation than of instruction or student ability. Salmon-Cox (1981) found that 50% of teachers surveyed used standardized test data only to confirm their own expectations: Teachers tended to pay attention to students' test scores only if the scores were higher than their own teacher-set marks. Kellaghan, Madaus and Airasian (1982) confirmed this finding in their study of testing in Ireland. Madaus (1981) claimed that teachers will modify their attitudes and behaviour only when test results are a key factor in deciding the life chances of individual students (e.g., grade promotion).

A second issue of concern to teachers is that of balancing testing time and teaching time. This appeared in the results of a survey of teachers' attitudes to testing (Herman and Dorr-Bremme, 1983). In this survey, the use of tests in 91 school districts in the United States was examined. The teachers reported that, although testing is a technique for motivating students to study harder, it also reduces the amount of time that can be spent teaching non-tested subjects or skills. A 1982 report from the Canadian Teachers' Federation (CTF) expresses a similar view and questions the benefit of testing programs on student learning.

A third issue concerns the use of test scores to evaluate teaching and teachers. The 1983 study of teachers' attitudes by Herman and Dorr-Bremme found that teachers did not want to be held accountable for students' test scores. Popham and Rankin (1981) also maintained that teachers do not want to be held responsible for pupil deficiencies that are due to external forces. The CTF Report (1982) cautions that tests should not be used to measure teacher effectiveness; it recommends that teachers be involved in establishing province-wide testing objectives.

To summarize:

a) Teachers may believe external exams provide an incentive for students to study (e.g., Yeh, 1978; Herman and Dorr-Bremme, 1983).

b) Teachers may feel that too much time will be spent preparing students to take the examination, time that will be lost to teaching (e.g., Canadian Teachers' Federation, 1982; Herman and Dorr-Bremme, 1983).

c) Teachers may resist the introduction of examinations because they fear being evaluated on the basis of the exam scores of their students (e.g., Popham and Rankin, 1981; Canadian Teachers' Federation, 1982).

3.3 Effects on Other Public Groups

In recent Alberta and U.S. surveys of public attitudes to examinations and students, the concept of public has been refined in different ways. Dumont (1977) isolated eight groups: general public, students, teachers, employers, principals, trustees, consultants, superintendents, and post-secondary deans and department heads. Ratsoy (1983) identified five groups: parents and other lay groups, in-school groups, school boards, Alberta associations, and post-secondary individuals as well as institutions. Gallup (1984) used three categories for the general public: no children in school, public school parents, and nonpublic school parents.

The Alberta surveys focused on perceptions of standards and reaction to the introduction of a testing program. Dumont (1977) found that the public was concerned by the fact that a common "yardstick" or standard had been lacking ever since the discontinuation of compulsory departmental exams in 1973. Public groups not involved in education made up about 50% of Dumont's sample, and they expressed a preference for a grading system involving a combination of teacher marks and departmental examination marks. A subsequent survey by Reid (1978) indicated that the public perceived standards to be declining. Ratsoy (1983) surveyed the public's reactions to Alberta Education's proposed introduction of compulsory comprehensive examinations for all graduating Grade 12 students. There was surprisingly weak support for compulsory examinations; only 21% of the respondents were in favour, and another 22% provided conditional support if certain aspects of the proposal were changed (e.g., make the examinations more like the former Departmental exams which were course-specific, and administer the exams to academic students only). The most frequently mentioned concerns included the negative impact of exams on dropouts and the non-academic student, the discrediting of school-awarded marks, and the narrowing of high school curricula and student choice of courses.

A public survey conducted by the Ontario Institute for Studies in Education (Livingstone and Hart, 1979) posed two questions about centralization in the organization of the school system. Thirty-five percent of respondents felt that there should be more control of the school curriculum at the provincial level, and 44% felt that provincial test results should be a criterion for judging progress in higher grades. The interpretation of these results is complicated by the fact that preference for centralization of curriculum development at the provincial level was not highly correlated with a preference for increased use of provincial tests (Livingstone and Hart, p. 20). Five annual surveys later, Livingstone, Hart and Davie (1985) concluded that "public opinion seems to be asking for greater external testing of student achievement" (p. 23). In the 1984 survey, 67% of respondents agreed that "province-wide testing should be used to assess individual performance of high school students" (1985, p. 26).

Several references to public opinion surveys were found in the American literature. Gallup polls, conducted annually for the past sixteen years, have measured the public's attitude toward the public schools. In 1983, 75% of respondents agreed that standardized national tests should be used to compare the achievements of students across the nation. This was the same response rate as that observed in 1970 (Elam, 1983). In 1984, 65% agreed that all high school students in the U.S. should be required to pass a standard nationwide examination in order to obtain a high school diploma (Gallup, 1984).

The introduction of examinations may reinforce the attitudes of various public groups, and allay or incite public concerns for:

a) the state of educational standards (e.g., Dumont, 1977; Reid, 1978);

b) the procedures whereby high school students are graduated (Gallup, 1984);

c) the use of exam results to compare students, schools and school systems (Gallup, 1984);

d) the impact of exams on lower achieving students (Ratsoy, 1983);

e) the distinction between central control of the curriculum and use of central examinations (Livingstone and Hart, 1980);

f) the narrowing of student choice (Ratsoy, 1983);

g) the narrowing of the curriculum (Ratsoy, 1983).

CHAPTER 4:
POTENTIAL EFFECTS ON THE TEACHING-LEARNING PROCESS

"The most prevalent danger of any...objective-based certification system...is the tendency to focus upon the starting and ending points of instruction with insufficient concern for the process of education" (Madaus and Airasian, 1977, p. 81).

In the above quotation, Madaus and Airasian identified the issue that external examinations might affect the teaching-learning process. An "ends-approach" to education will tend to emphasize the end product (the test score), as opposed to the process (the instructional activities). A goal-oriented approach to education may lead to neglect of the study of the instructional process itself.

4.1 Effects on Classroom Interactions and Activities

The following discussion includes attitudes of students and teachers in classrooms, as well as the pedagogical activities carried out in the classroom.

Examinations can be expected to have an impact on the classroom activities of students and teachers, including their interactions. Both groups involved in the teaching-learning process face the dilemma that under a system of external examinations they have two sets of learning objectives, not one: (i) the curriculum and (ii) test-taking strategies. These sets of objectives do not necessarily involve similar activities.

The activities of students may be motivated by the desire to beat the examination system, an objective that may be quite separate from learning the subject (Bloom, 1969). To accomplish this, cramming and rote memorization are common techniques, as noted by Entwistle (1981, p. 261). Sharp and Thomson (1984) have commented that examinations cause students to adopt an expedient approach to learning rather than to develop a commitment to learning.

Teachers will try to fulfil their responsibilities to promote learning and to prepare students for province wide examinations (Canadian Teachers' Federation, 1982; Herman and Dorr-Bremme, 1983). An additional responsibility, particularly for teachers of graduating students, is to provide their students with experience in the kinds of work and study habits required in post-secondary studies and examinations. Several writers have commented that exam preparation activities in secondary schools do not normally instill the type of independence and study habits that post-secondary institutions expect (Makins, 1977; Entwistle, 1981).

There is much speculation but little evidence about the changes that external exams cause in the classroom. Madaus and Airasian (1977) stated that exams have tended to determine the instructional emphasis when "they have some import for pupils and teachers" (p. 83). In particular, they suggested that external exams foster mechanization of the teaching and and learning process. Makins concluded, on the basis of an analysis of reports from several schools in London, England, that exams were causing narrow and didactic teaching, with more amphasis being placed on written competence, and less emphasis on oral and aural competence (1983).

The working relationship of students and teachers merits consideration. Under a system of external examinations, this relationship has been described as an alliance; they "work together at overcoming a 'common enemy' - the external examiners" (Makins, 1977, p. 3). But although both groups work together towards a common goal, they are not likely to be equal partners in the enterprise. It would seem that teachers will be very much the leaders, and students the followers. Makins referred to the teacher's control of students' working habits, with many small pieces of work assigned and done regularly. Reporting on the new Chinese examination system, Epstein (1982) noted that the exams have reinforced the distance between the roles of teacher and student, with teacher expertise becoming more highly valued and student dependency maintained.

In summary, the following effects of external examinations may be expected:

a) Teachers and students may be forced to alter their objectives and activities in the classroom (e.g., Entwistle, 1981; Canadian Teachers' Federation, 1982; Herman and Dorr-Bremme, 1983).

b) There may be increased camaraderie between the teacher and students, as they work toward a common goal (e.g., Makins, 1977).

c) Students may become more dependent on teachers for organizing their work, setting deadlines, etc., and this dependency may not hold them in good stead when they enter university (e.g., Makins, 1977; Entwistle, 1981).

4.2 Effects on the Implemented Curriculum

As well as to changing classroom interactions and activities, the introduction of examinations might also be expected to change the curriculum taught in the classroom.

4.2.1 Narrowing the curriculum of subjects with examinations

Recently published articles in the London Times Educational Supplement (e.g., Sayer, 1982; Makins, 1983) have criticized the external examination system in Britain for distorting, dominating and narrowing the curriculum to the point where the examination boards are alleged to be dictating curricular change. [This hypothesized effect of testing has not been substantiated by empirical evidence.]

Ratsoy's analysis of public responses to compulsory examinations in Alberta showed that a major public concern was that exams would result in an "exam-driven curriculum" (1983, p. 23). There were three aspects to this concern:

- Curricular emphasis might shift in the direction of objectives that could most easily be examined externally.

- The curriculum might become "watered down" in those examination subjects that all graduating students take because teaching will have to accommodate the least able students.

- The imposed curriculum may severely limit the optional courses students can take, thereby focusing the curriculum on examination subjects.

Madaus and Airasian (1977) state that "when there is a choice between emphasizing tested or nontested objectives, it is general experience that the objectives actually tested assume primacy" (p. 85). They go on to say that both teachers and students strive for the objectives made explicit in external exams rather than those made explicit in the curriculum guidelines.

The foregoing discussion suggests the following effects of external examinations:

a) Teachers will become more selective about curricular objectives (e.g., Ratsoy, 1983).

b) After the first examination year, teachers will choose objectives defined by examination questions (e.g., Madaus and Airasian, 1977).

4.2.2 Narrowing the curriculum of subjects with no examinations

"Whenever the outward standard of reality (examination results) has established itself at the expense of the inward, the ease with which worth (or what passes for such) can be measured is ever tending to become in itself the chief, if not sole, measure of worth. And in proportion as we tend to value the results of education for their measurableness, so we tend to undervalue and at last to ignore those results which are too intrinsically valuable to be measured." (Holmes, 1911, p. 128)

Non-tested subject areas have been affected by a narrowed curriculum, an effect which "is a consequence of the importance ascribed by society at large to test scores and of an emphasis on basic skills" (Herman and Dorr-Bremme, 1983, p. 15). The perception of "having a good education" may come to mean attaining a level of excellence in tested subjects rather than in non-tested subjects. When priorities change to emphasize test scores (and, thus, the subjects that are tested), the investments of time and energy on the part of students and teachers will also change. A Canadian Teachers' Federation report (1982) discussed the effects of exams on non-tested subjects: "If music and art are not subject to province-wide testing, does this mean they are not important?.... Will schools be forced to emphasize the teaching of skills for credentials over teaching for social competence?" (p. 2)

Effects on subjects with no examinations might include:

a) a lack of interest in these subjects (e.g., Herman and Dorr-Bremme, 1983);

b) less time for these subjects in the school timetable (e.g., Canadian Teachers' Federation, 1982).

CHAPTER 5:
POTENTIAL EFFECTS ON INSTITUTIONAL POLICIES

Perhaps because of the complexity of institutional policy-making and the problems of mounting empirical studies of institutional policy-making, no reports have been found that display data on this topic. Several discussion articles do touch on related issues, and these are summarized in this chapter.

5.1 Evaluation Policy

"Faced with the choice between having test results that are misused and having no test results, knowledgeable teachers will generally vote for no test results...There is danger that public pressure for publication of test results may destroy the very information teachers and administrators need to make wise decisions" (Coffman, 1980, p. 3).

A policy that may follow the introduction of province-wide examinations is the use of exam results in the evaluation of teachers, administrators and schools. Madaus and Airasian (1977) noted that "one hidden agenda in the competency-based (testing) approach is teacher accountability" (p. 88). This hidden agenda was also uncovered by Tyler et al.(1978) in their evaluation of the minimum competency testing program implemented in Florida; Tyler et al. described "the use of students' scores ...... as the major criterion for evaluating a teacher's effectiveness in the classroom" (p. 36). As mentioned in section 3.2, teachers do not want test scores used to measure their effectiveness; this reluctance is surely one reason for teacher resistance to the introduction of an examination system.

There are many examples of policy or the wish for policy whereby test scores would be used for evaluating teachers or schools. For example, Epstein (1982) reported that in China the reputations of schools and teachers are affected by the percentage of students who pass the exam and continue on to university. A recent survey of 196 Alberta school trustees by Webber (1984) showed that a majority of trustees want to use test results to compare teachers and schools.

Coffman made it clear that this would constitute a misuse of test results (1980). He noted that in any comparison of schools based on test scores, variation in school context (e.g., student abilities, family support of education, and student mobility) cannot be ignored. Gallup (1983) also discussed this problem.

> Earlier survey reports have pointed out that comparisons should take full account of the composition of the school population. Comparisons are only valid if the local school population reflects the national population. Schools that draw students from poor neighborhoods where parents have had little education and where language barriers exist obviously cannot be expected to achieve the same levels of test scores as schools in high-income communities (Gallup, 1983, p. 38).

It is interesting that these evaluation policies and intentions seem to have developed very shortly after the testing programs were introduced (e.g., the minimum competency movement in the late seventies in the U.S., the Chinese examination system in 1977, the Alberta comprehensive examinations in 1983). In fact, it is likely that there was not sufficient time to evaluate the testing programs on their merits before it was suggested that test results be used to evaluate groups or institutions within the educational system.

Examination effects on evaluation policy might include the following:

a) A policy of using exam scores to evaluate schools and teachers (e.g., Webber, 1984).

b) A policy introducing system-wide student evaluation procedures that include exam scores in the calculation of students' final grades (e.g., Tyler et al., 1978).

5.2 Post-Secondary Admissions Policy

Present admission procedures in Ontario universities are based almost exclusively on teacher marks; this is a matter of concern to many university educators (Traub and McLean, 1984). Local studies have attempted to relate success at university to Grade 13 marks (Etkin and Leathem, 1978; Traub, 1979) and to experimental admissions exams (Council of Ontario Universities, 1979). The Interface studies (Traub et al., 1977) found that the Grade 13 mark average correlated quite well with first year university mark average (a correlation of 0.64). If examinations are introduced in the graduating year of high school, they will almost certainly be used as a criterion in the universities' admissions process.

These considerations suggest the following effects of examinations on post-secondary admissions:

a) OAC exam scores are translated into admissions criteria by Ontario colleges and universities (e.g., Council of Ontario Universities, 1979).

b) The types of students admitted to selective university programs may change.

5.3 Communications Policy

Every phase of a testing program, from planning to interpretation of results, should be reported to a variety of audiences -- educators, students, legislators, trustees and the general public. One of the criticisms of the Florida minimum-competency testing program was the lack of adequate communication and the failure to involve the groups most seriously affected (teachers, students, and parents) in decisions. (Tyler et al., 1978, p. 31). If public perceptions of educational standards are important, then communication of accurate information is also important. Recent American reports have stressed the importance of communication about testing programs and the use of test scores (e.g., Roeber, 1980; Massachusetts State Dept. of Education, 1982).

Because of the existence of misinformation and speculation during the time when the reintroduction of exams in Ontario has been under consideration and because of the controversial nature of province-wide examinations, considerable effort will have to be made to convey correct information to the various audiences with a stake in the educational system.

Examinations may cause the following aspects of a communications policy to be debated and clarified by policy-makers:

a) the reasons for communicating information about exams, such as informing the general public, evaluating the exam program, promoting curriculum improvement, and providing help to decision makers (e.g, Roeber, 1980; Massachusetts State Dept. of Education, 1982);

b) the types of information to be communicated, such as the rationale for reinstituting exams, the curriculum objectives being examined, and the exam scores (e.g., Tyler, 1978; Coffman, 1980);

c) the sectors of the educational community to whom information will be communicated, such as the general public, school boards, schools, subject coordinators, teachers, and students (e.g., Tyler _et al_., 1978; Roeber, 1980).

CHAPTER 6:
CONCLUSIONS

6.1 General Conclusions

One conclusion of this review is that an impact study on the effects that the introduction of province-wide testing might have in Ontario would not duplicate research done elsewhere.

A second conclusion is that a great deal of caution must be exercised when one attempts to apply findings from other contexts and other geographic regions to Ontario. One subject of research and discussion papers has been the standardized, minimum competency testing movement in the U.S. This type of testing has few implications for an end-of-course examination system. One of the more important points drawn from the review of literature is that the effect of a testing program on individuals depends on the uses to be made of the test scores; this is also one of the major dangers in generalizing from existing research on the effects of a testing program. To identify the effects of an examination program it is first necessary to identify these uses.

6.2 Specific Conclusions

Discussions of the effects of examinations on the individuals directly concerned (i.e., students and teachers) pervade the literature. These effects vary from feelings of self-worth and fear of failure to activities in the classroom and students' future chances of success.

The effects of examinations on the teaching-learning process might be considered an "intrusion" on the classroom. Classroom teaching may become more expedient and tactical, and less discovery-oriented.

The effects of examinations on institutional policy may come later but the effects are likely to last longer. Policies concerning evaluation, admissions, and communications are most likely to change as a result of such innovations in the educational system.

6.3 Types of Effects

Four types of effects have been identified in Part One of this report, Strategies for Evaluating the Impact of Province-Wide Examinations. These types of effects are considered under the headings: perceptions, curriculum, enrolment trends and marking standards. The following notes are included to link this review to that part of the report.

The literature on perceptions covers a number of groups affected both directly and indirectly by examinations. Many writers have discussed the arousal of anxieties in students and teacher and changes in their attitudes. The attitudes of other groups towards standards and external examinations have also been of interest.

In general, the literature on curriculum effects has consisted of discussion papers rather than empirically-based reports. Part of the reason for this lamentable state of affairs has been the tendency of researchers to analyse the end result, the examination scores, rather than to focus on the curricular processes that led to the result.

Effects on enrolment trends have not been studied; however, much can be deduced from the literature on perceptions. Concern has been expressed over the problems facing the non-academic, non-tested student when there are external examinations. It is possible that the long-term effects on non-tested students may have consequences for enrolments in non-tested subjects.

The effects of examinations on marking standards have not been reported in the available literature. The reason, perhaps, is that most educational systems with external examinations have a long history of exams, and do not have a baseline for comparison. Ontario is well situated to collect baseline information before monitoring the effects of an examination program.

LIST OF REFERENCES

Airasian, P.W., and Madaus, G.F. "Linking Testing and Instruction: Policy Issues". Journal of Educational Measurement 20, (1983) pp. 103-118.

Airasian, P.W., Madaus, G.F., and Pedulla, J.J., eds. Minimal Competency Testing. Englewood Cliffs, New Jersey: Educational Technology, 1979.

Anderson, S.B., and Helmick, J.S., eds. On Educational Testing. San Francisco: Jossey-Bass, 1983.

Atkinson, D.R. "State-mandated Minimum Competency Testing Programs: Implications for School Counsellors". School Counsellor 29 (1981) pp. 22-27.

Barnette, J.J., and Thompson, J.C. III. "A Descriptive Assessment of the Effect of Evaluations on Instruction". Studies in Educational Evaluation 5, (1979) pp. 77-86.

Bloom, B.S. "Some Theoretical Issues Relating to Educational Evaluation". In Educational Evaluation: New Roles, New Means edited by R.W. Tyler, pp. 26-50. Chicago: University of Chicago Press, 1969.

Canadian Teachers' Federation. Province Wide Student Assessment Programs. Discussion Paper. Ottawa: Canadian Teachers' Federation, 1982.

Council of Ontario Universities. Experimental Achievement Testing Programme: Summary Report. Toronto: Council of Ontario Universities, 1979.

Coffman, W.E. "Those Achievement Tests - How Useful?". Iowa University, Institute for School Executives, 1980. (ERIC Document No. ED 209 762).

Crum, R., and Parikh, A. "Headmasters' Reports, Admissions and Academic Performance in Social Sciences". Educational Studies 9 (1983), pp. 169-184.

Dumont, F.J. Alberta Grade 12 Examination Study: A Study Commissioned by the Minister's Advisory Committee on Student Achievement (MACOSA). Edmonton: Province of Alberta, 1977.

Dumont, F.J. Alberta Grade 12 Examination Study: Condensed Version. A MACOSA Study. Edmonton: Province of Alberta, 1977.

Dumont, F.J. Alberta Grade 12 Examination Study: Executive Summary. A MACOSA Study. Edmonton: Province of Alberta, 1977.

Dunn, S.S. Public Examinations: The Changing Scene. Adelaide: Rigby, 1974.

Dunn, T.R. "An Empirical Demonstration of Bias in HSC Examination Results". The Australian Journal of Education 26 (1982), pp. 190-203.

Elam, S.M. "The Gallup Education Surveys: Impressions of a Poll Watcher". Phi Delta Kappan 65 (1983), pp. 26-28.

Entwistle, N. Styles of Learning and Teaching. Chichester: John Wiley, 1981.

Epstein, I. "An Analysis of the Chinese National Examination: The Politics of Curricular Change". Peabody Journal of Education 59 (1982), pp. 180-189.

Etkin, B., and Leathem, B. Grade 13 Marks as a Predictor of Performance in Engineering (Parts I and II). Toronto: University of Toronto, 1978.

Gallup, G.H. "The 15th Annual Gallup Poll of the Public's Attitudes Toward the Public Schools". Phi Delta Kappan 65 (1983), pp. 33-47.

________. "The 16th Annual Gallup Poll of the Public's Attitudes Toward the Public Schools". Phi Delta Kappan 66 (1984), pp. 23-38.

Glossop, J.A., and Roberts, C. "An Exploratory Study of Examination Policy Differences and Performance in Three Comprehensive Schools". Educational Review 32 (1980), pp. 67-85.

Gray, J. "A Competitive Edge: Examination Results and the Probable Limits of Secondary School Effectiveness". Educational Review 33 (1981), pp. 25-35.

Haertel, E., and Calfee, R. "School Achievement: Thinking About What to Test". Journal of Educational Measurement 20 (1983), pp. 119-132.

Hatala, R.J. "Testing in Perspective". New Directions for Testing and Measurement 16 (1982), pp. 141-145.

Herman, J.L., and Dorr-Bremme, D.W. "Uses of Testing in the Schools: A National Profile". New Directions for Testing and Measurement 19 (1983), pp. 7-17.

Homes, E.G.A. What Is and What Might Be: A Study of Education in General and Elementary in Particular. London: Constable, 1911.

Kellaghan, T.; Madaus, G.F.; and Airasian, P.W. The Effects of Standardized Testing. Boston: Kluwer-Nijhoff, 1982.

Lewis, D.M. "Certifying Functional Literacy: Competency Testing and Implications for Due Process and Equal Educational Opportunity". Journal of Law and Education 8 (1979), pp. 145-183.

Ligon, G.D. "Preparing Students for Standardized Testing". New Directions for Testing and Measurement 19 (1983), pp. 19-27.

Livingstone, D.W., and Hart, D.J. Public Attitudes Toward Education in Ontario: 1979. Toronto: OISE Press, 1980.

Livingstone, D.W.; Hart D.J.; and Davie, L.E. Public Attitudes Toward Education in Ontario: 1984. Toronto: OISE Press, 1985.

Madaus, G.F. "Testing and Funding: Measurement and Policy Issues". New Directions for Testing and Measurement 1 (1979), pp. 53-61.

__________. "Reactions to the Pittsburgh Papers". Phi Delta Kappan 62 (1981), pp. 634-636.

__________, and Airasian, P.W. "Issues in Evaluating Student Outcomes in Competency-based Graduation Programs". Journal of Research and Development in Education 10 (1977), pp. 79-91.

__________, Kellaghan, T.; Rakow, E.A.; and King, D.J. "The Sensitivity of Measures of School Effectiveness". Harvard Educational Review 49 (1979), pp. 207-230.

__________, and McDonagh, J.T. "Minimum Competency Testing: Unexamined Assumptions and Unexplored Negative Outcomes". New Directions for Testing and Measurement 3 (1979), pp. 1-14.

Makins, V. "Why Cream of Sixth Goes Sour". The Times Educational Supplement 3243 July 29, 1977, p. 3.

__________. "Exams Cause of Narrow Teaching". The Times Educational Supplement 3491 May 27, 1983, p. 12.

Massachusetts State Department of Education. "Interpreting and Using Commercial Achievement Test Results. Basic Skills Improvement Policy: Supplement to Implementation Guide I". Boston, 1982. (ERIC Document No. ED 221 589).

McClung, M.S. "Are Competency Testing Programs Fair? Legal?". Phi Delta Kappan 59 (1978), pp. 397-400.

Moore, W.E. "Some Functions of Examinations". In Public Examinations: The Changing Scene edited by S.S. Dunn, pp. 51-75. Adelaide: Rigby, 1973.

Passmore, B. "Welsh Pupils More Reluctant to Stay on for CSE Exams". The Times Educational Supplement 3436 May 7, 1982, p. 10.

________. "Public Exams - Main Cause of Welsh Under-Achievement". The Times Educational Supplement 3485 April 15, 1983, p. 14.

Popham, W.J., and Lindheim, E. "Implications of a Landmark Ruling on Florida's Minimum Competency Test". Phi Delta Kappan 63 (1981), pp. 18-22.

________, and Rankin, S.C. "Minimum Competency Tests Spur Instructional Improvement". Phi Delta Kappan 62 (1981) pp. 637-639.

Ratsoy, E.W. Public Reactions to the Proposed Provincial Student Evaluation Policy. Edmonton: Alberta Education, 1983.

Reid, J.E. "Inflation of Standards: Fact of Fiction?". A paper presented at the Annual Meeting of the Canadian Educational Researchers Association, London, Ontario, 1978.

Roeber, E.D. "Teaching Local Educators to Use and Report State Assessment Results". Michigan, 1980. (ERIC Document No. ED 211 570).

Salmon-Cox, L. "Teachers and Standardized Achievement Tests: What's Really Happening? Phi Delta Kappan 62 (1981), pp. 631-634.

Sarason, I.G. "Understanding and Modifying Test Anxiety". In On Educational Testing edited by E.B. Anderson and J.S. Helmick, pp. 133-149. San Francisco: Jossey-Bass, 1983.

Sayer, J. "Why Profiles are More Attractive". The Times Educational Supplement 3442, June 18, 1982, p. 4.

Serow, R.C., and Davies, J.J. "Resources and Outcomes of Minimum Competency Testing as Measures of Equality of Educational Opportunity". American Educational Research Journal 19 (1982), pp. 529-539.

Sharp, A., and Thomson, G. "Performance in External Examinations and Pupils' Orientations to Studying". Educational Review 36 (1984), pp. 37-51.

Sproull, L., and Zubrow, D. "Standardized Testing from the Administrative Perspective". Phi Delta Kappan 62 (1981), pp. 628-631.

Traub, R.E. "Unsupported and Iniquitous: A Proposal by Bernard Etkin and Brian Leathem". Commentary prepared for the Research and Evaluation Branch, Ontario Ministry of Education, 1979.

________, and McLean, L.D. "A Rosy View -- University Admission Officers' Preferences and Expectations for Provincial Examinations". The Ontario Institute for Studies in Education, 1984. (unpublished report)

________; Wolfe, R.; Wolfe, C.; Evans, P.; and Russell, H.H. Secondary-Postsecondary Interface Project II: Nature of Students. Volumes I and II. Toronto: Ministry of Education and the Ministry of Colleges and Universities, Ontario, 1977.

Trusz, A.R., and Parks-Trusz, S.L. "The Social Consequences of Minimum Competence Testing". Educational Studies 12 (1981), pp. 231-241.

Tyler, R.W. Educational Evaluation: New Roles, New Means. The Sixty-eighth Yearbook of the National Society for the Study of Education, Part II. Chicago: University of Chicago Press, 1969.

Tyler, R.W. et al. "Impact of Minimum Competency Testing in Florida". Today's Education 67 (1978), pp. 30-38.

Webber, C.F. "School Board Member Perceptions of the Utility and Importance of Student Evaluation Information in Alberta". Planning Services, Alberta Education, 1984. (unpublished report)

Wigdor, A.K., and Garner, W.R. Ability Testing: Uses, Consequences, and Controversies. Parts I and II. Washington, D.C.: National Academic Press, 1982.

Yeh, J.P. "Test Use in Schools: Studies in Measurement and Methodology, Work Unit 4". Center for the Study of Evaluation, Los Angeles, 1978. (ERIC Document No. ED 214 951).

SECTION II:
The Impact of Assessments

CHAPTER 1:
AN INTRODUCTION

Before discussing the potential impact of a provincial assessment of education, it is necessary to acknowledge the problem of definition. Assessment has been defined and therefore interpreted in different ways. The writings of Bloom (1970), Satterly (1981) and Wood (1984) reveal this confusion in terminology.

Bloom (1970) described assessment as only one of three aspects of the testing enterprise, the other two being measurement and evaluation. By assessment Bloom meant "attempts to assess the characteristics of individuals in relation to a particular environment, task, or criterion situation" (1970, p. 30). He emphasized the fact that assessment ought to be as much concerned with the environment as with the individuals who interact with the environment.

Satterly offered a general definition of assessment:

> Educational assessment is an omnibus term which includes all the processes and products which describe the nature and extent of children's learning, its degree of correspondence with the aims and objectives of teaching and its relationship with the environments which are designed to facilitate learning (1981, p. 2).

Satterly also used the terms 'impressions' and 'constructions' to distinguish between informal and formal assessment. The first referred to observations of the student's performance recalled later, the second to more deliberate assessment procedures conducted with established criteria.

Wood (1984) discussed the variation among working definitions of assessment; he expressed the need for educational researchers to reconceptualize the term 'assessment', and to separate the term from 'measurement' once and for all.

Designers of an assessment program for Ontario will need to achieve consensus on a definition of terms. Attention to such details can ensure that few problems of misunderstanding the intent of an assessment, or its monitoring, occur.

1.1 An Overview of Implemented Assessment Programs

In the literature, assessment programs were described at three levels of implementation: international, national, and intranational. Most references to the impact of an assessment were reports of, or reactions to, one of three assessment programs -- the IEA, NAEP and APU. The first of these is an international association for educational assessment, while the others are national assessments, one for the United States and the other for England, Wales and Northern Ireland. All three programs are described in more detail in the following sections, which are organized by level of implementation.

1.1.1 International assessment

The International Association for the Evaluation of Educational Achievement (IEA) has existed for over twenty-five years. It "attempts to establish a science of empirical comparative education based on close cooperation between institutions in many countries" (Husen, 1979, p. 371). The first survey, a cross-national examination of cognitive development in children, was proposed in 1958, and led to a 12-country study of mathematics achievement. It was followed by a Six Subject Survey of 21 countries, including four developing countries (Passow, Noah, Eckstein and Mallea, 1976). IEA studies have recently been conducted in secondary mathematics and science. Ontario and British Columbia were each treated as "countries" in the mathematics study, and nine Canadian provinces (Quebec excepted) participated in the science study.

Later sections of this paper will refer to lessons learned from the early IEA studies, because these suggest possible effects in Ontario (Husen, 1979; Thiesen, Achola and Boakari, 1983).

1.1.2 National assessment

Power and Wood (1984) reviewed and compared three national programs designed to "define, assess, and monitor student achievement at a national level" (p. 355). These are the American National Assessment of Educational Progress (NAEP), the British Assessment of Performance Unit (APU), and the Australian Studies in Student Performance (ASSP).

1.1.2.1 The National Assessment of Educational Progress (NAEP)

The NAEP was designed in the late 1960's as a survey testing program. It was intended to monitor performance at national and regional levels, not at state and school-district levels. The program assesses performance and monitors changes in achievements of students in the grade that is usual for nine, thirteen and seventeen year olds in ten subjects, including art, citizenship, literature, mathematics, reading, science, social studies, and writing (Greenbaum, Garet and Solomon, 1977).

More recently, there has been a shift in control of the structure of NAEP from the Education Commission of the States (ECS) to the Educational Testing Service (ETS). There has been an accompanying shift in emphasis from that of dissemination of regional information to that of assisting individual states to adapt the national assessment model.

1.1.2.2 The Assessment of Performance Unit (APU)

The APU was set up in 1975 to promote methods of assessing and monitoring student achievement at a time when standards were believed to be deteriorating. Two initial programs, consisting of five consecutive years of annual monitoring of mathematics (11 & 15 year old students) and science (11, 13 & 15 year old students), have been completed. The intention is to replace annual monitoring with periodic monitoring at five year intervals (Gipps and Goldstein, 1983).

1.1.2.3 Australian Studies in Student Performance (ASSP)

The ASSP began in 1974 because of concern about the problems of disadvantaged groups. This led the Australian Council for Educational Research (ACER) to undertake a national survey of basic skill levels in literacy and numeracy for a sample of 10 and 14 year olds. As a result of the 1975 survey, the government pressed for the establishment of a national system for monitoring standards. Finally, in 1979, the Australian Studies in Student Performance (ASSP) was initiated to provide national data on performance in basic skills. ASSP was permitted to release only minimal findings: national breakdowns by year (1975 vs. 1980), sex, and location (urban vs. rural). The original intent had been to collect assessment data for five consecutive years. It was later decided not to proceed with further testing, due in part to changes in the political context, opposition from the teachers' unions, and the minimal results released for the 1980 study (Power & Wood, 1984, p. 359).

1.1.3 Intranational assessment

1.1.3.1 Canadian provincial assessments

Several Canadian provinces conduct assessments (Canadian Teachers' Federation, January 1980; McLean, 1982). They differ in types of items or instruments, subject areas, grade levels and sampling procedures. Some provinces use standardized tests (e.g., Nova Scotia uses both the Thorndike Intelligence Test and the Metropolitan Achievement Test). Other provinces have developed or are developing item pools with the assistance of teachers and subject matter experts. One example of this is the Ontario Assessment Instrument Pool or OAIP. While items of the multiple-choice variety predominate, most provinces have included other forms of items (McLean, 1982, p. 80).

The subject areas tested include mathematics, science, language arts (or reading and writing), and social studies. Each subject in a given grade is assessed on a 3,4, or 5 year cycle. Several sequences of grade levels are assessed (e.g., 3,6 & 9 in Alberta, 4,7 & 10 in British Columbia, and 9 & 12 in Nova Scotia). Sampling procedures range from every-student testing in British Columbia to random sampling in Manitoba.

Provision for optional testing of students from outside the sample means that the assessments can serve an evaluation function at other levels. For example, Manitoba teachers can opt to have all students in their class write the tests, while Alberta school boards can opt to have additional schools included in the assessment.

1.1.3.2 Other intranational assessments

An objective of the NAEP is to assist individual states to apply national assessment technology (Greenbaum, Garet and Solomon, 1977, p. 12). Sebring and Boruch (1983) reported on the number of states which have recently begun to use this opportunity. In addition, there are many assessment models designed by individual states. One example is the California Assessment Program (CAP).

One of the objectives of the APU was to promote assessments in cooperation with Local Education Authorities (LEAs) (Hextall, 1984, p. 245).

1.2 An Advance Organizer

The preceding discussion has been a brief introduction to the assessment programs reported in the literature. The rest of the paper consists of five chapters devoted to different aspects of assessment impact.

Chapter 2, on the consequences of assessment planning decisions, acknowledges three components of an assessment that may contribute to its effects. These are the purposes of the assessment, the nature of the contextual measures, and the technical specifications.

Chapter 3 outlines some potential effects on individual students, teachers and other public groups, as reported in the literature.

Chapter 4 is a discussion of potential effects on the teaching-learning process, specifically of effects on teaching, the implemented curriculum and evaluation of the teaching-learning process.

Chapter 5, on the potential effects on institutional policy, covers curriculum, assessment and communications policies.

Chapter 6 contains some conclusions about the potential effects of an assessment program.

CHAPTER 2:
IMPLICATIONS OF PLANNING DECISIONS

The discussion in this chapter assumes that the various components of an assessment program each contribute to the impact of the program in unique ways. The following three components are considered:

1. the purposes of the assessment;

2. the nature of the contextual measures in the assessment;

3. the technical specifications of the assessmment model.

2.1 Purposes of Assessment

The literature on the impact of assessments cannot be understood without giving due consideration to the purposes of each assessment. The effect of the assessments has been monitored (at least in part) through an evaluation of the extent to which the purposes of the assessment were achieved. The following purposes of assessment might be expected to produce different effects on the educational system:

- Monitoring the system: the assessment serves to produce an objective statement of student achievement.

- Conducting research on the system: ancillary data are used to form variables for students, classrooms or schools; these variables are correlated with, and, with supporting rationale, may be presumed to cause, student achievement.

- Evaluating the system: the assessment data are interpreted to yield a statement about the quality of student achievement (good, bad, indifferent) and the quality of instructional programs, of schools and school systems.

In a review of the NAEP, APU and ASSP programs, Power and Wood (1984) found that each program was primarily conceived as a monitoring project rather than a research project. "Each project has tried to distinguish between describing the current status and changes in performance and explaining the how and why of observed levels, variations, and trends in performance" (p. 364).

Although the initial emphasis of national assessments appeared to be on monitoring as opposed to research, Power and Wood felt that this appearance did not jibe with the appearance suggested by later developments in the three programs (Power and Wood, 1984). For example, although the public pronouncements of the APU ignored research and stressed monitoring, Gipps and Goldstein (1983) noted that traditional research issues predominated at all stages, "from initial discussion of items to writing of final reports" (p. 157). The conclusion of Gipps and Goldstein was that "if anything of real use is to emerge then it will do so as the result of a high-quality research effort rather than a narrowly conceived monitoring exercise" (1983, p. 164).

Husen (1979) criticized the early IEA studies for their reliance on "an input-output model [of education, their breadth] of scope, and [their] emphasis on quantitative methods and statistical techniques with no reliance at all on qualitative observations and anthropological methods" (p. 384). Although the early IEA studies were pioneering ventures, Husen maintained that the limitations of the research paradigm had a lot to do with the weaknesses of the research.

> The straightforward paradigm with representative samples and strict quantitative and standardized methods to test hypotheses uniformly over a number of age levels and countries seemed at the time to be self-evident. We never seriously considered an alternative strategy, for example, limiting ourselves to a selection of a few schools and classrooms that could be subjected to intensive, qualitative observations. We certainly expected too much from the broadly collected information that was obtained by questionnaires from the students about their home background and from the teachers about how they taught (Husen, 1979, p. 382).

Wood and Power (1984) stressed the importance of including classroom measures based on observational data in assessments:

> From now on, any project which aims to survey what schools manage to do with students ... ought to concern itself, as best it can, with the whole business of schooling and, above all, with teaching and learning (Wood and Power, 1984, p. 319).

It would seem that the experienced designers of previous large-scale assessment paradigms have a lesson for proponents of assessments in the 1980's: There is a need to make assessment results more directly relevant to classroom practices and, thus, to approach the realities of teaching and learning more closely (McLean, 1982, p. 95).

Adopting a certain purpose for the assessment may have the following consequences:

a) The purpose of monitoring might yield objective statements about achievement, as reflected in item responses; but it might not provide the correlational information needed to explain and understand the response data.

b) The purpose of research might be expensive in time and expertise, but it might also generate answers to questions about teaching and learning.

c) **The purpose of system evaluation might imply comparisons of different classes, schools and boards and, thereby, guarantee a lack of cooperation from teachers and other educational groups.**

2.2 Contextual Measures

The term 'contextual' is used here to denote measures of student environment and background. Both the IEA and the national assessments have been criticised for failing to obtain adequate information about contextual variables. The shortcomings of the IEA studies in this regard have been discussed by Thiesen, Achola and Boakari (1983, p. 46). These shortcomings include the failure to collect data related to the social, demographic, and environmental characteristics associated with school settings. For example, the IEA Second International Mathematics Study (SIMS) included only one question about the social environment of the school (whether it was in a rural or urban environment), and only four questions about student background (age, sex, parental education and occu-pational status).

> Conspicuously missing are items dealing with school selectivity, general level of district resources, local occupational opportunities, socioeconomic status of local residents, school learning environment, or related indicators of economic/cultural context (Thiesen, Achola and Boakari, 1983, p. 47).

To enhance the interpretation of national data, Thiesen et al. (1983, p. 67) suggested that the following seven clusters of variables be measured:

1. occupational aspirations of the student -- both level and type;

2. general educational and occupational aspirations and expectations prevalent in the local environment;

3. job opportunities and remuneration in occupations related to different disciplines;

4. quantity and emphasis of instruction that occur within schools, especially in cases where samples stratified by regions, SES (socio-economic status), and so forth are likely to produce sharp cross-sectional differences in these measures;

5. classroom environment measures such as the extent to which independence is fostered, authority is exhibited, encouragement is supplied, and so forth;

6. perceived importance of the subject by the students as assessed by:

   a) level of parental encouragement;

   b) importance of achievement in the subject to the student's status in society;

c) relationship of achievement in the subject to educational and occupational goals;

7. general structure of the educational system and the opportunity and values inherent in it (Thiesen et al., 1983, p. 47).

With the exception of the fifth and seventh clusters of variables, these would be easy to include in questionnaires for students or teachers.

The shortcomings of the national assessment programs in obtaining information about context have also been noted. Power and Wood (1984) described the limitations of the NAEP, APU and ASSP. Power and Wood noted the failure of these programs to measure variables that would permit an investigation of the relationship between performance and environmental characteristics. This failure was felt to have been a direct result of conceiving the purpose of assessment as monitoring rather than conducting research. Thus:

> inadequate provision [had been made] for the difficulties of interpreting performance levels in the absence of home and school background and process data or the additional information needed if the results were to be linked with other educational and social data, so as to inform policy and deepen our understanding of what is happening in the nation's schools (p. 364).

Consequently, the results were difficult to interpret and "in the absence of other data and a research component, have contributed little to policy or to public understanding" (p. 365). Power and Wood suggested that data from all three assessments would have been more useful if "additional student and school background and process data [had been] collected and further research on the instruments and follow-up studies [had been] undertaken" (p. 376).

Greenbaum, Garet and Solomon (1977) discussed the failure of several contemporary research efforts to identify explanatory variables of achievement (p. 107). They claimed that the NAEP had not made a systematic attempt to find out which background variables would be most appropriate to measure in relation to the academic achievement of students. Responding to this criticism, the NAEP claimed to have funded a review of literature on the association between educational outcomes and background variables, and to be considering the implications of this review for NAEP policy. The object of this was to assist NAEP in presenting better descriptive data and to "strengthen the analysis and the interpretations" of achievement data (Greenbaum, Garet and Solomon, 1977, p. 209).

In their treatise on the APU, Gipps and Goldstein (1983) reported that the APU has begun to look at different background measures in trying to account for student performance. Many of the original background measures used (e.g., pupil/teacher ratio and region of the country) had been of little interest to policy makers, school districts and teachers.

> More relevant variables which would relate to the circumstances in which children learn, for example, size of teaching group, qualifications and experience of the teacher resources available (particularly for science) and aims of

> the programme of work, have been used in the later surveys (1983, p. 161).

Gipps and Goldstein expected that forthcoming reports would either confirm the utility of these variables or recommend that this type of information be collected more effectively through in-depth studies rather than through large-scale surveys.

The foregoing discussion of contextual variables suggests the following matters for consideration in designing an assessment:

a) Relevant contextual measures may require other methods of data collection than questionnaires.

b) Non-traditional research methods have high costs for data collection and analysis, which might prevent their inclusion in an assessment program.

c) Case studies might be used to collect information on contextual measures in a less costly way.

2.3 Technical Issues of Assessment

The technical specifications of an assessment affect its impact. Hextall (1984) referred to technical specifications as the "cloak of technicism and expertise" (p. 260) that the public assumes to be mystical in nature. He argued that technical decision making should not be left to experts:

> The establishment of assessment principles, criteria for evaluation, and testing procedures is a crucial social process which must be rendered more open to scrutiny, not least by those who are to be the subject of assessment. This demand implies much fuller public and collective debate of the basic ground-rules upon which any mode of assessment is formulated (Hextall, 1984, p. 260).

Four technical issues will be discussed in the following sections:

1. sampling procedures
2. the nature of the assessment instruments
3. reporting procedures
4. the testing cycle

2.3.1 Sampling procedures

The way in which examinees and test items are sampled seems likely to affect the impact of an assessment. If sampling were not used and every student responded to the same items, the greatest impact would likely be on individuals (as in an examination system). However, this would mean that coverage of different curricular objectives, and hence the ability of the assessment to reflect the whole of the curriculum, would be severely restricted. On the other hand, matrix sampling from a large pool of test items could represent the curriculum better, but have little effect on individuals.

The assessment instrument can be an evaluative tool for the classroom teacher, particularly when every student in the class has taken the test and the results are available to the teacher. The results might be used to find areas that require remedial treatment. If the item sampling procedure included the same core of items on every form of the instrument or if every student in a class were to take the same form, the results might be included in the grade of students who participated in the assessment. This result might also be achieved if teachers were encouraged to use the assessment instruments in subsequent student evaluation.

Husen (1979), in discussing the technical problems of the early IEA surveys of achievement, mentioned proper sampling as a problem that surfaced from the beginning of the studies. Until that time, few countries had had experience with drawing representative, random samples of school and student populations for educational research (p. 375).

Hextall (1984) described the sampling process used in an APU assessment of science. The science team had conducted a number of surveys using a 1.5 percent sampling of all eligible students. In the 1980 study of 11 year olds, this meant surveying 11,000 students in 1097 schools out of a possible 800,000 students in the same age cohort in 19,362 schools. Six categories of scientific performance were identified for testing, none of which was administered to as many as a third of the sample of students. By averaging, it turned out that results for two of the six categories of performance were based on only 1 of 200 and 1 of 600 students respectively, or 1 student in every 5 schools and 1 student in 15 schools respectively. Despite this limited sample, results for science achievement were publicized in the press, with the following generalizations:

- "Too many dunces in science class";
- "Science is all fun but no method to haphazard 11 year olds";
- "Boys do better than girls in primary science";
- "Science: Sex differences start young, APU reports show" (Gipps and Goldstein, 1983, Appendix 9).

Hextall (1984) believed that this type of analysis raised questions "about the claim of the APU to be collecting national facts and figures on which to base significant statements about standards in schools" (Hextall, 1984, p. 253).

McLean (1982) discussed how poor sampling procedures (of students and behavior) can threaten the validity of the assessment. He found that most Canadian provincial assessment reports ask the reader to accept the quality of the sample of students on faith alone.

> Given that the essence of assessment is generalization to the student population of the province, and given that designing such a sample requires not only high technical skill but also a good knowledge of schools and access to detailed information (school types and sizes, community characteristics, etc.) both the designed and achieved samples deserve more attention (p. 90).

2.3.2 The nature of the assessment items

The items selected for inclusion in the assessment instruments can affect the impact of the assessment program. Two reasons are discussed below:

- public access to the items

- the difficulty level of the items

2.3.2.1 Public access to items

If test items are available to teachers prior to administration of the assessment instrument, then it is possible that the items will have an impact before the assessment is conducted. We can imagine teachers using the instruments in their own teaching to give students practice responding to the items. If test items (and corresponding results) are available to teachers after the assessment has been completed, then teachers might use this information to assess the performance of their students and evaluate the success of their own teaching strategies.

The NAEP has had a policy of releasing about 50 percent of test items after each assessment. Until now, the items used in APU assessments have been unavailable to teachers. Gipps and Goldstein (1983) reported that once the current five-year program of annual monitoring is over, the APU plans to make the items available. This is in line with their perception of the future role of the APU: "concentrate on making more use of the information that is available by opening the item banks, making data available to interested researchers, and improving dissemination among teachers" (p. 159).

The implication of item availability is that there is a fairly large pool or bank of items from which test items will be drawn for each assessment. Along with matrix sampling, Shoemaker

(1976) believed that item banking provided the essential technology for "placing educational assessment on a firm foundation" (p. 226).

There are two conflicting views of item pools or item banks. One is that a pool or bank is simply a collection of test questions; the other is that the test questions are calibrated, that is response data are added to the pool or bank so that examiners can find out how well examinees have answered particular items.

2.3.2.2 Item difficulty

A measure of how well (or how poorly) students have answered a particular test item can be calculated as an index of difficulty. The difficulty of items constituting an assessment instrument affect the assessment results, and therefore, the assessment's impact. Power and Wood (1984) emphasized the importance of this for the results of the national assessment programs they reviewed.

> In the long run, the results one gets from national testing depend very much on the difficulty of the items utilized. The three approaches (of the national assessments) lead to the development of tests that differ markedly in the difficulties of the items employed, and this leads in turn to quite different pictures of national performance. It is not easy to recognize how much the image, created by national assessment, of the health of the education system depends on the test development strategies employed .... the outcome of a national assessment is a function of the characteristics of the instruments used (Power and Wood, 1984, p. 366).

To illustrate this, contrast the items selected by the NAEP and APU. For the NAEP, the objective was threefold: to assess what most, what typical, and what relatively few students could do. Thus, equal numbers of easy, average and difficult items were included in the assessment tests. In contrast, the objective of the APU was to assess the student norm, and items of 50 percent or median difficulty were thought to have been favoured for the assessment tests (Power and Wood, 1984, p. 366).

2.3.3 Nature of the reporting

The purpose of an assessment determines to some extent the nature of the reporting. Consider the different types of reports that might be generated by assessments which adopt the three different purposes of monitoring the system, conducting research, and evaluating the system. If the purpose were to monitor the system, then reports would likely consist of item-level results, with little in the way of interpretative information. If the purpose were to conduct research, then reports would likely reflect the interests and specializations of the researchers who chose the ancillary variables. If the purposes included evaluating the system, then reports would contain conclusions based on interpretations of the data and reflecting the value systems of the evaluators.

Power and Wood saw the goal of the three national assessment programs as having been that of monitoring; this was reflected in the types of reports that were published: neutral, non-interpretive, facts-only. The lack of interpretation was advanced as the reason why the assessments were of limited use to decision makers and practitioners. At the same time, Power and Wood remarked that the dull, low-key quality of reports ensured they did not upset people (if, indeed, they were read at all); this may have accounted for the survival of at least the APU in its early days (p. 371).

It may be a fact of program development that reporting becomes an issue in the later stages of a program. For example, the early years of NAEP were consumed with other issues:

> the problems of contract monitoring, sampling, data analysis, and objectives and exercise development; little time, money, or staff were available for consideration of NAEP's end product, its reports (Greenbaum, Garet and Solomon, 1977, p. 213).

Later on, NAEP did realize the necessity of getting information in "selectively focused" (p. 214) ways to educators, legislators, and lay audiences.

McLean (1982), writing about Canadian provincial assessments, commented on the absence of any concrete plans to feed results back to schools. "The implication seems to be that every-one will know what to do when they receive the monitoring results" (p. 94). He cited British Columbia as the only province with a systematic process for discussing results and possible follow-up actions at the local level (p. 95).

Dissemination of findings among teachers is a recent priority of the APU. To this end, they now organize regional conferences, publish a series of occasional papers, and publish newsletters. However, "much of the value of the APU's results will be lost without better dissemination. What is required is a means of informing teachers and others and discussing with them the implications of the APU's work within the context of the curriculum and different methods of assessment" (Gipps and Goldstein, 1983, p. 165).

Barnes, Moriarty and Murphy (1982) advanced the view that achievement testing is a dangerous activity, one reason being that results are often misinterpreted or exaggerated by various groups. "In fact, the public typically views the results of district testing as reflective of the value, quality, or respective effectiveness of a school or school system" (p. 14). For this reason, it is important to interpret test results for each different consumer group: board of education members, central administration, students, parents, teachers, counselors, and the press.

2.3.4 The testing cycle

The testing cycle refers to the testing interval, or the frequency with which assessment tests are administered. Several programs have reported frequency of testing as a problem that may affect the impact of an assessment for two reasons. One reason is that, if the period between assessments of the same subject is short, only minor changes in performance are likely. Greenbaum, Garet and Solomon (1977) recommended that NAEP consider the cost-benefit of widening the testing cycle to ten years if it were found that most of the changes being measured in the second cycle were minor ones.

A second reason that the frequency of testing has been reported as a problem is the fact that it may leave insufficient time for detailed analysis, interpretation and reporting before the next round of data collection. Greenbaum, Garet and Solomon (1977) suggested that increasing the length of time between tests might result in improvements in the quality of testing instruments and interpretations (p. 177). Another example is the APU's decision to go from annual monitoring to a five-year cycle. Gipps and Goldstein (1983) expressed the belief that this change would give the research team the needed time to explore the data, carry out in-depth studies and develop efficient survey designs (p. 166).

Husen (1979) also noted problems that arise when time limitations restrict the analysis of assessment information. In the IEA studies, according to Husen, the technical personnel "who assisted us in planning data processing and statistical analyses were steering us, not we them. The data sets, therefore, were highly under-analyzed" (Husen, 1979, p. 383).

2.3.5 Implications of technical choices

Choices on technical matters might have the following effects:

a) The sampling procedures might guide the overall assessment design and, in an ad hoc way, eventually determine the extent to which achievement is reported.

b) Making test items available to teachers before the assessment might cause teachers to reflect on their own expectations of student knowledge and integration of curriculum material, and might also cause teachers to compare the assessment objectives with their own objectives.

c) Reporting only item-level data might limit the impact of the assessment on educators due to the time they would have to spend studying the results; choosing to provide a condensed report at the meta-objective level might limit the impact if the results were less easily interpreted.

d) A short testing cycle may mean that the assessment data is not fully analyzed, while an extremely long testing cycle might render meaningless any comparisons of performance across the time interval between two testings.

CHAPTER 3:
POTENTIAL EFFECTS ON INDIVIDUALS

Very few reports have dealt with the effects of large-scale academic assessment on individuals. In part, the reason may be that assessment programs have tended to be less concerned with measuring individuals, and more with measuring the educational system. The technical decisions in assessments are taken in the knowledge that it is group responses rather than individual responses that are of interest. The use of multiple matrix sampling procedures means that the cohort of examinees responding to each item is a representative sample of eligible participants, and that different cohorts respond to different assessment instruments. Also, only a sample of classes and schools might be included in an assessment. Thus, an assessment is intended to provide a snapshot of the way in which the educational system is performing.

This chapter is a review of potential effects of assessment on three types of individuals:

1. students

2. teachers

3. other public groups

3.1 Effects on Students

Little research has been uncovered on the effect of assessment on individual students. This might mean that assessments have a minimal effect on individual students. The only identified effect has been on attitudes to test-taking.

Assessment programs seem not to motivate students to perform well on the assessment tests. The relatively high frequency of items omitted (particularly of items requiring open-ended answers) in the OAIP field trials supports the conclusion that assessment instruments may be assessing motivation or willingness to cooperate as much as, or even more than, knowledge of the subject.

Omvig (1971) reviewed the literature on student motivation and concluded that students tend to be careless and unmotivated in their performance on tests unless they are personally concerned about their own test scores. Omvig was particularly interested in the problem of student motivation on standardized achievement tests used in studies of the effects of school size or class size. According to Omvig, "[m]ore often than not, the students are involved only as the producers of test results" (p. 47), which are subsequently used to answer research questions. He questioned whether "the students [were] sufficiently motivated to put forth the effort required to insure valid test results" (p. 47).

In Omvig's experimental study, an attempt was made to solve the problem of student apathy towards standardized tests. He designed a "pre-test" session in which a school counsellor discussed with individual students their past standardized test results, drawing attention to those areas in which poor achievement had been displayed and praising the student for the progress made in other areas. It was hypothesized that this "treatment" would produce more valid standardized test scores. However, the results for 270 ninth grade students did not support this hypothesis.

It would appear that the problem of student apathy has not been been solved. In Madaus' (1981) study on the impact of standardized testing, he found that behaviours and attitudes towards testing were not likely to change if the tests did not affect the individual's life chances. Bearing this in mind, one might make the assessment more important to students by providing feedback on test results to examinees and their teachers. This would be particularly useful if every eligible student in every class, school and board had been tested. However, this type of thinking emphasizes the individual rather than the group response, and clouds the distinction between assessment and examination.

In the absence of evidence, the only recourse in speculating about the effects of an assessment on students is to common-sense. Some potential effects include the following:

a) If the assessment is perceived to be meaningless and unrelated to their studies, then students are likely to be unmotivated in responding to the test items.

b) If the pool of assessment items is sufficiently large so that typical items are made available to teachers and students prior to testing, then students can better understand what is expected of them, and they will be more highly motivated as a result.

c) If the assessment results will be used to evaluate students, classes or schools, and if students know this, then they may be more highly motivated to do well on the test.

d) If individual students receive feedback on their performance soon after being tested, then they may be more positively motivated.

3.2 Effects on Teachers

According to the literature, assessments have caused teachers to react in one of two ways: by expressing interest in helping plan the assessment, and by expressing concern that assessment results might be used to evaluate teachers.

3.2.1 Teacher participation

In the United Kingdom, the teachers' unions were invited to join a consultative committee for development of the APU. Although the teachers acted as a constraining force by limiting the number of background variables measured, limiting the study of personal and social development, and eliminating the study of ethnic groups, any concern about such effects as opposition of the teaching force to assessment, was dissipated by the involvement of the teachers (Gipps and Goldstein, 1983, p. 48). In fact, results of a national survey indicated that 70% of headteachers were in favour of the APU, 18% were neutral, and 12% opposed (Gipps, Steadman, Blackstone & Stierer, 1983, p. 120).

In Australia, the teachers' unions were not invited to join the steering committee for the 1980 ASSP assessment, so teachers strongly opposed the program, and many boycotted it. As a result, 22 percent of participating schools withdrew too late to be replaced. According to Power and Wood (1984), the failure to get the teachers' support proved fatal to the entire Australian assessment program (p. 364).

3.2.2 Monitoring programs or monitoring teachers?

A primary concern of teachers has been that assessment results might be used to evaluate their teaching, thereby monitoring teachers in addition to programs. Power and Wood (1984) acknowledged the importance of responding to this concern in setting up an assessment program:

> Attempts to set up systems designed to assess the performance of public institutions have invariably met with resistance from those whose work is to be monitored. The major political problem faced in establishing a national assessment program has been to gain the cooperation of the teaching profession (Power and Wood, 1984, p. 362).

For example, it had been a major task for the developers of NAEP to convince "teachers, principals, and district superintendents that the NAEP could not and would not be used to evaluate the performance of individual teachers, schools, programs, districts, or even states" (Power and Wood, 1984, p. 362).

It may be that for teachers, the proof is in the pudding; positive experience with an assessment program will yield positive attitudes towards assessment. Wood and Gipps (1982) reported on a British assessment program that included a sampling design, yet gave schools the opportunity to test students who were not part of the sample. In one school district, with a 10% sampling policy, 92% of schools tested 100% of eligible students. Wood and Gipps maintained that this would not have been possible had teachers not begun to trust the authority enough to ask for every-student testing.

The foregoing discussion suggests these effects on teachers:

a) If the pool of items is large enough that the test items are released to teachers prior to the assessment, then teachers may learn what the expectations of student performance are.

b) If feedback is supplied to the participating teachers about the performance of their students, then teachers may implement new teaching strategies in cases where objectives had been taught but not learned by the students.

c) If assessment results are perceived as a means of evaluating teachers, then teachers may cooperate minimally, or not at all, with the assessment program.

3.3 Effects on Other Public Groups

Assessment programs, such as the APU and ASSP have been initiated at times when public concern about the quality of education led to the feeling that educational standards were deteriorating. In considering the effects of assessment on public perceptions of standards, one can concentrate on two phases of an assessment: the introduction of the program, and the reporting of results. The former may be sufficient to change perceptions of educational standards, and the latter, to give the impression of educational accountability. Each is discussed in turn.

3.3.1 Perceptions of educational standards

> National assessment programmes, whether or not they mean to, promote a view of 'standards' (that is to say, a view of what education is meant to be doing) which, in its emphasis on basic or minimal accomplishments of a severely selected kind, is narrowing and limiting, and definitely not conducive to the emergence of flexible and imaginative educational policies designed to cope with the future (Wood and Power, 1984, p. 319).

3.3.1.1 Defining the term 'Standards'

'Standards' is an unclear term in the educational literature, used frequently with the adjectives low and poor. The implication of the resulting phrase is that the educational system is not operating as well as it should. Despite the popularity of the term in journal articles, as well as in the rationale for the need for assessment programs, the term has not been well defined.

Wood and Power (1984) noted that a national assessment could contribute to clarifying the meaning of 'standards' both as defined and reported. Lapointe and Koffler (1982) perceived the role of the NAEP in the search for standards as one of measuring student achievement and reporting it publicly. But the search continues for working definitions of 'standards' and 'higher educational standards':

> A 'higher' or 'better' educational standard is one that measures and reports on what are critical elements of desired student achievement. The demand [for standards] refers not only to higher student performance, but also to the inclusion

> of broader curricular coverage and attention to higher order skills (Lapointe and Koffler, 1982, p. 5).

3.3.1.2 Baselines as standards

According to Power and Wood (1984), all three national programs (i.e., APU, ASSP and NAEP) have chosen to keep out of the standards debate and let others use the assessment results to form their own impressions of how well the school system is functioning. For these programs, standards of performance referred to the bench marks provided by the assessment results: therefore, they were relative or pseudo standards rather than absolute standards.

> Standards in the absolute sense remain on the agenda; national assessment programs provide indirect, equivocal, and circumstantial evidence of changes in pseudo standards, which is not clear or interpretable enough to enrich our understanding of what is happening in schools (p. 376).

Gipps and Goldstein (1983) discussed the APU's original intention to identify and define the standards of performance students might be expected to achieve. However, this attempt was soon translated into "describing measured performance over a period of years - a less contentious task" (Gipps & Goldstein, 1983). Thus, for the APU, standard came to mean a baseline measure comprised of composite measures produced by five years of monitoring.

3.3.2 Perceptions of accountability

One effect of an assessment may be satisfying the need for accountability in education. For example, Hextall (1984) discussed the possible use of APU results in meeting the accountability needs of Local Education Authorities (LEAs); their needs include:

- the need to evaluate the quality of schools and teaching across the region;

- the need to have an objective basis for allocating dwindling resources.

Originally, the APU rejected the notion that nationwide monitoring might be used as a basis for educational accountability, but later papers underlined the connection between assessment, accountability and resource allocation (Hextall, 1984, p. 255). At a time of severe cutbacks in education expenditures in the United Kingdom, test results could offer LEAs a "valuable means of judging the cost-effectiveness of schools, departments, or methods of organizing curricula and teaching" (p. 256).

Wood and Gipps (1982) described two LEAs that introduced testing programs for the purpose of allocating resources. Although the LEAs maintained that information gained from testing was only one factor in decision making, Wood and Gipps believed that the test results would influence any decision, since there is "a strong tendency for quantitative data to overwhelm other sources of information, whatever the protestations to the contrary" (p. 53).

3.3.3 Potential effects on public perceptions

The effects of an assessment on public perceptions of standards might include the following:

a) If an assessment were initiated, then the public might think that something was being done to increase standards.

b) Assessment results might confirm existing perceptions that the educational system is not succeeding, or they might suggest that the system is in better shape than previously thought.

c) Comparing assessment results over two testing cycles might lead to the conclusion that educational standards are either improving, declining or staying about the same.

CHAPTER 4:
POTENTIAL EFFECTS ON THE TEACHING-LEARNING PROCESS

The present chapter considers assessment impact on the teaching-learning process. Teaching and learning are considered to be interactive processes in the classroom and will be discussed as one activity. Literature on impact on teaching-learning has concentrated on the teaching part of the process. This is not surprising considering that the literature on effects on individuals (see Chapter 3) has tended to concentrate much more on effects on teachers than on effects on students.

This chapter is divided into three effect areas:

1. teaching

2. the implemented curriculum

3. evaluation of the teaching-learning process

4.1 Effects on Teaching

Most of the literature reported in this chapter and accepted as evidence of contributions by assessments to teaching methodology has resulted from NAEP assessments. The research reported in many of the articles was funded by the National Science Foundation (NSF) of the United States and the National Council of Teachers of Mathematics. This suggests that the professional organizations have been eager to use the NAEP data.

4.1.1 Teaching practices

The test items and results from an assessment can be used to guide instructional practice. Sebring and Boruch (1983) conducted an exploratory study on the uses made of NAEP results and found that most results were used in professional ways, "employing NAEP data, methods, and materials to improve educational research programs and instruction" (p. 17). Teachers have been encouraged to use the NAEP results in two ways. First, they were encouraged to examine the NAEP results for particular content objectives (especially those results that were poor) and consider whether their teaching strategies had contributed to the poor results. Second, teachers were encouraged to use the published NAEP exercises (and corresponding results) to assess the performance of their students in relation to national or regional results.

Hiebert (1981) reported on the contribution that NAEP results made to knowledge about elementary students' conceptions of unit of measure. He described students' responses to several exercises on the NAEP mathematics assessment. These provided insight into the lack of student understanding of basic properties of units. For example, they "do not fully understand that a unit of measure may represent more than a single entity" (p. 38). Then, he discussed the implications of the percentage response rates for the NAEP sample and concluded that:

> it is not appropriate to think of children as completely possessing or lacking a measurement concept, but rather as being able to apply the understanding they have in particular situations. Some children demonstrate knowledge of a concept in simple tasks but appear to abandon this knowledge in more complex settings (p. 42).

In his recommendations for instruction, he suggested that teachers administer the NAEP exercises reproduced in the article in order to confirm their own students' degree of understanding of the topic.

Lindquist, Carpenter, Silver and Matthews (1983) compared the results of the second and third NAEP assessments of mathematics for elementary and middle schools. A major finding was that the significant gains that were noted were due to improved performance of routine exercises, such as computation, and that students had made no gains on exercises assessing deep understanding or applications of mathematics. Lindquist et al. elaborated on this finding in a consideration of results for whole numbers, fractions and decimals, and other basic concepts and skills. The article included NAEP exercises (as well as national percentiles) which were used to interpret performance in these content areas. Then the authors posed the following questions about teaching techniques:

- Does class discussion focus on the variety of interpretations or representations that might be possible, or do students see only a single solution for a problem?
- Are students asked to defend their reasoning, or justify an answer, or explain why a particular result is reasonable?

4.1.1.1 Teaching to the test

An assessment can influence what is taught in classrooms by focusing on the subjects and objectives being tested. Wood and Power (1984) discussed the possibility of this effect in American states that have state-wide assessment programs, particularly in cases where assessment results are perceived to be a measure of the quality of a school's program. An easy way to raise scores (hence, an easy way to substantiate the claim to have raised 'standards') would be for schools to teach to the assessment tests. Wood and Power quote newspaper reports from California describing how schools in San Diego County registered sizeable score increases on the annual state assessment tests (California Assessment Program, or CAP) by remodelling their curriculum to emphasize material tested by the CAP, and by teaching test-taking skills (p. 316).

Another example of teaching to the test was cited by Gipps and Goldstein (1983). Head teachers of some primary schools in a particular LEA insisted that reading be treated as a subject when it was discovered that the LEA had introduced a reading testing program (p. 189).

4.1.1.2 Teaching tools

The effect of an assessment on teaching tools, such as textbooks, in classroom activities has been suggested in the literature as a secondary effect of assessment; but no empirical evidence has been uncovered.

Forbes (1977) concluded that the NAEP, which was designed as a long-term project whose major impact would come only after it had reassessed several learning areas, had already proven useful to many audiences; one such audience was professional educators who had interpreted the NAEP results in terms of their implications for textbook improvements. Textbook coverage of the objectives tested by the NAEP has been found inadequate in many instances. Hiebert (1981) concluded that students needed much more experience with situational problems on the topic of unit of measurement and that the kinds of experiences and activities needed had not been included in most textbooks. Thus, "it is important for the teacher to supplement the textbook program with measuring activities ... which are designed to facilitate children's understanding of the basic unit measurement concepts" (p. 43).

Lindquist, Carpenter, Silver and Matthews (1983) discussed the more general problems of students' failure to gain a deep understanding of mathematics. Their questions about the quality of textbooks included the following:

- Do textbooks place sufficient stress on the higher-level objectives, or do they dwell too heavily on routine knowledge?

- What supplemental materials are available to extend the text in the direction of higher-level objectives?

4.1.2 Potential effects on teaching

The foregoing discussion on the effects of assessment on teaching practices suggests the following effects:

a) Individual teachers might evaluate their own teaching objectives and content coverage by examining the published samples of provincial assessment items and objectives.

b) As a group, teachers might use assessment results to assess how effective their teaching strategies have been in relation to student performance on assessment items and objectives.

c) In response to exceptional student responses (good and bad) to assessment items, the teaching profession might suggest new strategies to produce better results.

d) Textbooks and other instructional aids might become targets of the concerns of educational groups about poor assessment results, especially on curriculum objectives deemed to be important.

4.2 Effects on the Implemented Curriculum

Gipps and Goldstein (1983) believed that it was too soon to make a pronouncement about the impact of the APU on the curriculum. They did say, however, that fears about assessments dominating the curriculum and causing revisions have not yet been realized (p. 159). They attributed this lack of effect to two features of APU assessments: the sampling procedures and the lack of intelligible public reports. They predicted that curriculum might be affected by the nature of the assessment. If the assessment were subject-specific (e.g., limited to mathematics, language, science and modern languages), then the curriculum might become narrower and additional emphasis might be placed on these subjects. However, if the assessment cut across all curricular areas, it might shape the content of the curriculum differently. Gipps and Goldstein reported circumstantial evidence suggesting an effect of APU assessments may be to widen what is taught in the tested areas (1983, p. 159).

> It may well be that the APU's assessment material, capable in many cases of promoting a widening of the curriculum, will come to be seen as its greatest achievement, rather than any direct contribution to the debate on standards or accountability (p. 162).

Effects on the implemented curriculum might include the following:

a) If the assessment plan is to cover the entire curriculum, that is all subject areas and objectives, then assessment may broaden the curriculum coverage.

b) If the assessment plan is to cover a few subject areas and specified topics, then the impact may be a narrowing of the curriculum.

4.3 Effects on Evaluation of the Teaching-Learning Process

Large-scale assessments have had an effect on evaluation methodology: they have identified evaluation problems, instrumentation requirements, and so forth, specific to large, quantitative studies, and have attempted to resolve such problems and requirements through the development of statistical procedures. To illustrate this type of effect, the following discussion will consider two methodological problems common to large-scale assessments:

1. problems with measuring change in student achievement

2. problems with synthesizing assessment results

4.3.1 Measuring change in achievement

A goal of politicians, if not of the educators responsible for an assessment, is to track academic achievement over time. Husen (1979) criticised the design of the first set of IEA studies for not employing a longitudinal design, whereby the same students would be followed for at least one year and their gains in achievement during that time measured and studied (pp. 383-384). Thiesen, Achola and Boakari (1983) also believed that the second set of IEA studies should have collected longitudinal information over time. However, they acknowledged "the great distance [the IEA] knowledge base has advanced in the past 15 years" (p. 52). In other words, the technology available fifteen years ago may not have supported a longitudinal study.

Goldstein (1983) has recently worked on conceptual problems associated with measuring trends over time. He reported that, despite the fact that the APU and the NAEP gave a high priority to making inferences about trends over time, little attempt appears to have yet been made to define the meaning of trends over time or to discuss associated measurement problems. Using the APU for illustrative purposes, Goldstein (1983) outlined the most commonly used methods for measuring absolute change in achievement over time, and presented alternative formulations. He suggested that measuring relative changes over time, using either standardized differences or longitudinal analyses at the level of the school, would be feasible and could yield interesting results.

> If we discover that regional differences have narrowed and that this continues to remain the case even after a number of possible confounding variables have been allowed for, then we may have begun to uncover something interesting and useful (Goldstein, 1983, p. 377).

4.3.2 Synthesizing assessment results

Bock, Mislevy and Woodson (1982) speculated about the next stage in educational assessment. They claimed that the assessment movement had been shaped, if not actually made possible, by several developments:

1. the accountability movement

2. survey sampling techniques

3. matrix and multiple-matrix sampling techniques

4. creation of the NAEP

Bock et al. held that the next stage of growth would be in the area of reporting assessment results. The original method of reporting was described by the authors as a fixed-item approach, typical of social survey research. The idea was to report the percentage of correct responses to each item as well as the change in this percentage from one assessment to another.

The authors proposed instead that the random-item concept be applied to assessment results (p. 6). In this approach, the data are reduced to a small number of broadly interpretable attainment indices that reflect performance on sets of items sampled from specified domains by pupils sampled from specified populations. The use of item response theory to do this was described (p. 8).

The foregoing discussion suggests that assessment might affect evaluation of the teaching-learning process by stimulating:

a) the development of procedures for tracking trends over time;

b) the use of procedures involving more data reduction or synthesis so that item indices are replaced by achievement indices.

CHAPTER 5:
POTENTIAL EFFECTS ON INSTITUTIONAL POLICIES

Potential uses of assessment results by policy makers have been mentioned in the literature, although few examples of policy-related effects have been cited (e.g., Sebring and Boruch, 1983). That assessments have had little effect on institutional policies might be due to the fact that the purpose of assessments has usually been monitoring the system, not conducting research. Gipps and Goldstein (1983) believed that this restricted focus limited the usefulness of the APU for policy making and that if the APU were to become useful to policy makers, then "a high-quality research effort [would have to be mounted] rather than a narrowly conceived monitoring exercise" (p. 164).

A second reason why assessments may have had little or no effect on policy is the length of time from the inception of an assessment program to the dissemination of results. The problem of time is exacerbated by the fact that the initial cycle of an assessment program is often only a pilot, the results of which are used to modify and improve the design of the program.

The following sections on potential institutional policy effects cover the areas of curriculum, assessment and communications policies.

5.1 Effects on Curriculum Policy

Assessment may affect devleopment and evaluation of curriculum policy. In the initial stage of the assessment program, curriculum specialists and educators are forced first to determine and agree on the curricular objectives for testing, and then to assess the relevance of particular instruments and items for measuring those objectives. Later, the assessment results can be used to decide whether the curricular objectives had been met, and if not, to guide the modification of existing curricular policies. Kearney (1983) mentioned the use of assessment results to identify relative strengths and weaknesses within an educational system (or subsystem, such as a particular school within a district), and thereby to identify the system needs for resources.

Sebring and Boruch (1983) provided instances of boards and teachers using NAEP items to evaluate and modify local curriculum policies. This tended to happen more often when the district chose to conduct its own assessment: that is, the state tests were administered to a sufficient number of students to permit fair comparisons with state data (p. 17). It was reported that the state assessment for Minnesota had a positive impact on decisions about curriculum content and planning at the district level (p. 18). Also, seven state case studies revealed that local schools used NAEP resources to establish objectives and develop curricula as well as to engage in curriculum assessment and evaluation.

In one program, planned but not yet implemented, assessment results were intended to have an effect on policy for teacher training. This assessment model, developed in 1982 for Trinidad and Tobago, was described by Wood and Power (1984). The objective of the first round of testing was to expose deficiencies in mathematics attainment of primary school leavers. Once identified, these problem areas were to be addressed via changes in the teacher training program. A later round of testing primary school leavers was to check whether the changes in teacher training had improved achievement of primary school leavers (Wood and Power, 1984, p. 316).

Assessment effects on curriculum policy might include the following:

a) In planning an assessment, curriculum committees may need to refine and order their objectives, evaluate their curriculum guidelines, and thereby, revise curriculum policy.

b) In examining assessment results, curriculum committees may need to compare the relative achievement of students on different objectives and modify curriculum priorities accordingly.

5.2 Effects on Assessment Policy

An intended impact of several large-scale assessment programs has been to make other systems and subsystems more conscious of the potential uses of assessment activities for their purposes. For example, the APU and NAEP have included among their objectives the advancement and dissemination of assessment technology. In fact, the NAEP shifted focus from collecting and analyzing achievement data at national and regional levels to disseminating information about conducting assessments. They provided their model for the design and administration of state and local assessment programs, "encouraged the use of its released exercise items, and assisted states in designing tests tailored to their assessment needs" (Power and Wood, 1984, p. 374). By 1983, at least 12 states had copied the NAEP model and 14 had adapted the model (Sebring and Boruch, 1983, p. 17).

These observations suggest the following possible effects of provincial assessment on board assessment policy:

a) A school board might adopt or adapt a provincial assessment model for its own needs.

b) In adapting a provincial assessment model, a school board might choose to test every student and use the assessment results to evaluate students, classes, or schools within the board.

5.3 Effects on Communications Policy

In a discussion of the misuse of assessment results, Kearney (1983) referred to two communications policy issues. He believed that assessment planners need to find an effective means to communicate the purposes and disseminate the results of an assessment to policy makers. In addition, Kearney argued that policy makers at local, state and, possibly, national levels ought to be communicating the results of an assessment to the general public (p. 10). Evidence that communications policy is important to the success of an assessment program was presented in Chapter 2, under the heading, Nature of the Reporting.

Assessment planners may need to consider the following in developing a communications policy:

a) the potential uses for which assessment information can be communicated;

b) the types of information it will be necessary to communicate;

c) the potential audiences who can use the information in a positive way.

CHAPTER 6:
CONCLUSIONS

6.1 General Conclusions

One conclusion of the literature review is that few empirical studies of the effects of assessments on educational systems have been made. Assessment of student achievement is a recent phenomenon; it originated within the past twenty-five years on national and international educational levels. The early assessment programs can be considered exploratory, pioneering ventures, which had the effect of fostering advances in assessment methodology. It may be that insufficient time has elapsed since the beginning of national and international assessments for their impacts to have been seriously considered.

A second conclusion of this review is that the few examples of impact that are reported may not be applicable to other assessment situations because the effects have been confounded by such factors as the historical and political context, and by design specifications that are unique to particular assessments.

6.2 Specific Conclusions

The following specific conclusions summarize the findings of the present paper regarding three areas of potential effect -- individuals, the teaching-learning process, and institutional policies.

Effects on individuals directly involved in the assessment seem likely to be dependent upon such technical matters as the procedure used for sampling students and test items (curriculum). If an assessment were designed so that comparisons of individual students (or teachers) could be made, then the effect of the assessment on individuals might be increased. However, if comparisons of individuals were not possible, or, at least, not practical, then the assessment's effect on the individual would be less. Effects on other public groups might include changes in perceptions of educational standards and accountability, both through the introduction of an assessment and through the reporting of assessment results.

Effects on the teaching-learning process can be expected to take time. Before teachers and other educators can begin to interpret the results and use them in improving teaching strategies and changing teaching priorities, assessment results need to be analyzed and reported, possibly over several testing cycles.

Effects on institutional policies may occur in the areas of curriculum, assessment, and communications. Assessment may affect the planning, modification and evaluation of curriculum policies. Assessment policies at sub-system levels may incorporate the policies of large-scale assessment programs that encourage such adaptation.

LIST OF REFERENCES

Barnes, R.E.; Moriarty, K.; and Murphy, J. "Reporting Testing Results: The Missing Key in Most Testing Programs". National Association of Secondary School Principals' Bulletin 66 (1982), pp. 14-20.

Bloom, B. "Toward a Theory of Testing Which Includes Measurement-Evaluation-Assessment". In The Evaluation of Instruction edited by M.C. Wittrock and D.W. Wiley, pp. 25-50. New York: Holt, Rinehart and Winston, 1970.

Bock, R.D.; Mislevy, R.; and Woodson, C.. "The Next Stage in Educational Assessment". Educational Researcher 11 (1982), pp. 4-11.

Broadfoot, P. Selection, Certification and Control. London: The Falmer Press, 1984.

Canadian Teachers' Federation. Province-wide Student Assessment Programs - The Teachers' Response. Winnipeg: Canadian Teachers' Federation, 1980.

Forbes, R.H. "NAEP: One Tool to Improve Instruction". Educational Leadership 34 (1977), pp. 276-281.

Gipps, C., and Goldstein, H. Monitoring Children: An Evaluation of the Assessment of Performance Unit. London: Heinemann Educational Books, 1983.

Goldstein, H. "Measuring Changes in Educational Attainment Over Time: Problems and Possibilities". Journal of Educational Measurement 20 (1983), pp. 369-378.

Greenbaum, W.; Garet, M.S.; and Solomon, E.R. Measuring Educational Progress: A Study of the National Assessment. New York: McGraw-Hill, 1977.

de Gruijter, D.N.M., and van der Kamp, L.J.T., eds. Advances in Psychological and Educational Measurement. London: Wiley & Sons, 1976.

Hextall, I. "Rendering Accounts: A Critical Analysis of the APU". In Selection, Certification and Control edited by D. Broadfoot, pp. 245-262. London: The Falmer Press, 1984.

Hiebert, J. "Units of Measure: Results and Implications from National Assessment". Arithmetic Teacher 28 (1981), pp. 28-43.

Husen, T. "An International Research Venture in Retrospect: The IEA Surveys". Comparative Education Review 23 (1979), pp. 371-385.

Johnson, G.H. "Making the Data Work". Compact 6 (1972), pp. 29-30.

Kearney, C.P. "Uses and Abuses of Assessment and Evaluation Data by Policymakers". Educational Measurement: Issues and Practice 2 (1983), pp. 9-12, 17.

Lapointe, A.E., and Koffler, S.L. "Your Standards or Mine: The Case for the National Assessment of Educational Progress". Educational Researcher 11 (1982), pp. 4-9.

Lindquist, M.M.; Carpenter, T.P.; Silver, E.A.; and Matthews, W. "The Third National Mathematics Assessment: Results and Implications for Elementary and Middle Schools". Arithmetic Teacher 31 (1983), pp. 14-19.

Madaus, G.F. "Reactions to the Pittsburgh Papers". Phi Delta Kappan 62 (1981), pp. 634-636.

McCormick, R.; Bynner, J.; Clift, P.; James, M.; and Brown, C.M., eds. Calling Education to Account. London: Heinemann Educational Books, 1982.

McLean, L. "Educational Assessment in the Canadian Provinces". In Assessing Educational Achievement, edited by D.L. Nuttall. Special issue of Educational Analysis 4 (1982), pp. 79-96.

Minister's Advisory Committee on Student Achievement. Student Achievement in Alberta. Edmonton: Alberta Education, 1979.

Nuttall, D.L., ed. Assessing Educational Achievement. Special issue of Educational Analysis 4 (1982).

Omvig, C.P. "Effects of Guidance on the Results of Standardized Achievement Testing". Measurement and Evaluation in Guidance 4 (1971), pp. 47-51.

Ontario Ministry of Education, Research and Information Branch. The Ontario Assessment Instrument Pool: A Curriculum-based Aid to Evaluation. Review and Evaluation Bulletin, Vol. 1, No. 1. Toronto: Ministry of Education, 1979.

Passow, A.H.; Noah, H.J.; Eckstein, M.A.; and Mallea, J.R. The National Case Study: An Empirical Comparative Study of Twenty-one Educational Systems. New York: John Wiley, 1976.

Power, C., and Wood, R. "National Assessment: A Review of Programs in Australia, the United Kingdom, and the United States". Comparative Education Review 28 (1984), pp. 355-377.

Satterly, D. Assessment in Schools. Oxford: Blackwell, 1981.

Saylor, G. "How to Use the Findings from National Assessment (NAEP)". Education Digest 40 (1974), pp. 42-45.

Sebring, P.A., and Boruch, R.F. "How is National Assessment of Educational Progress Used?". Educational Measurement: Issues and Practice 2 (1983), pp. 16-20.

Shoemaker, D.M. "Applicability of Item Banking and Matrix Sampling to Educational Assessment". In Advances in Psychological and Educational Measurement, edited by D.N.M. de Gruitjer and L.J.T. van der Kamp, pp. 225-231. London: Wiley, 1975.

Theisen, G.L.; Anchola, P.P.W.; and Boakari, F.M. "The Underachievement of Cross-national Studies of Achievement". Comparative Education Review 27 (1983), pp. 46-68.

Tyler, R.W. "Educational Assessment, Standards, and Quality: Can We Have One Without the Others?". Educational Measurement: Issues and Practice 2 (1983), pp. 14-15, 21-23.

Womer, F.B., and Mastie, M.M. "Can National Assessment Change American Education?". Compact 6 (1972), pp. 26-28.

Wood, R. "Assessment Has Too Many Meanings and the One I Think We Want Isn't Clear Enough Yet". Educational Measurement: Issues and Practice 3 (1982), pp. 5-7.

__________, and Gipps, C. "An Enquiry into the Use of Test Results for Accountability Purposes". In Calling Education to Account, edited by R. McCormick et al., pp. 44-54. London: Heinemann Educational Books, 1982.

__________, and Power, C. "Have National Assessments Made Us Any Wiser about 'Standards'?". Comparative Education 20 (1984), pp. 307-321.

__________, and __________. Review of: W. Wirt & A.E. Lapointe "Measuring the Quality of Education: A Report on Assessing Educational Progress". Journal of Educational Measurement 21 (1984), pp. 209-212.

Wittrock, M.C., and Wiley, D.E. The Evaluation of Instruction: Issues and Problems. New York: Holt, Rinehart and Winston, 1970.

APPENDICES FOR PART THREE

A Review of Literature on Examinations and Assessments

Appendix A:
A Selectively Annotated Bibliography on the Impact of Examinations

THE IMPACT OF EXAMINATIONS: AN ANNOTATED BIBLIOGRAPHY

A. Journal Articles

B. Books, Published Reports and Unpublished Documents

C. Newspaper Articles

D. Documents from the Educational Resources Information Center (ERIC)

A. Journal Articles

Airasian, P.W., and Madaus, G.F. "Linking Testing and Instruction: Policy Issues". Journal of Educational Measurement 20 (1983), pp. 103-118.

Discusses Minimum Competency Testing (MCT) and some problems associated with content, curricular and instructional validity. Also refers to the legal implications and an article by M.S. McClung (1978), annotated later in this section.

Atkinson, D.R. "State-mandated Minimum Competency Testing Programs: Implications for School Counsellors". School Counsellor 29 (1981), pp. 22-27.

Discusses two possible effects of testing programs: test anxiety (p. 23) and failure (p. 25), with the need for counselling in both cases.

Barnette, J.J., and Thompson, J.C. III. "A Descriptive Assessment of the Effect of Evaluations on Instruction". Studies In Educational Evaluation 5 (1979), pp. 77-86.

A study of 224 secondary schools in a Northeast U.S. state. According to teachers' perceptions of evaluation, student performance affected instructional practice much more than either program or teacher evaluation. While observations of students was most often used (49%), standardized tests counted for only 5% of instructional change, which included reviewing, modifying and revising instruction to meet student needs.

Boreham, N.C. "An Evaluation of a Method of Monitoring Grade Standards in Examinations". Research In Education (Manchester) 22 (1979), pp. 74-85.

This was a U.K. study in which a statistical method for external testing was evaluated. It was found that one third of student grades could have been misgraded. (An external examiner re-marked 20 teacher-marked tests, and regraded them if the average grade assigned by a particular teacher was one-half grade out.) Later newspaper article by Doe (1980) refers to Boreham. Likely of limited interest to our study.

Brown, S., and McIntyre, D. "Influences upon Teachers' Attitudes to Different Types of Innovation: A Study of Scottish Integrated Science". *Curriculum Inquiry* 12 (1982), pp. 35-51.

This was a study of the implementation of two different types of innovations, Organizational and Pedagogical. Innovations of the Organizational type were presented clearly and carried out without any teacher control. However, the pedagogical innovations involved a change of teacher behaviour in the classroom, and conception of the changes was not imposed by an authority. There was no link between teaching activities and course objectives, and no explicit criteria for pupil performance. "The innovation will be implemented in any classroom only insofar as the individual teacher has a favorable attitude toward it, has the motivation, skills and resources to modify his current patterns of teaching, and understands what is meant by the innovation and how to go about introducing it." (p. 43)

Crum, R., and Parikh, A. "Headmasters' Reports, Admissions and Academic Performance in Social Sciences". *Educational Studies* 9 (1983), pp. 169-184.

This is a U.K. study of 158 university students in the social sciences; exam grades, qualitative attributes, and success in university were studied. Crum and Parikh recommended that universities attend to the headmaster's report on performance (standard of work) and recommendation (head's prediction of the success and potential a student possesses) as important indicators of degree performance.

Dixon, N.R. "Testing - Its Impact on Expectations, Practice, Accountability". *Educational Leadership* 35 (1978), pp. 294-297.

It is argued that standardized tests have too much power: use of such tests *shapes educational policy*, affects the *what* (acquisition of facts important) and *how* of teaching. With such tests, *electives are de-emphasized* and *the back to basics movement is emphasized*. In addition, teacher effectiveness and school quality are measured by test performance. "Another impact.... is that it has led to labelling and tracking students in the schools and in society. These tests have led to punitive educational discrimination for blacks, other ethnics, and the children of the poor" (p. 296). In terms of accountability, the author states that tests *assess narrow intellectual functions*, and omit so much that is critically important in both school and life.

The author's comments are based on reflection only, and not empirical evidence.

Dunn, T.R. "An Empirical Demonstration of Bias in HSC Examination Results". *The Australian Journal of Education* 26 (1982), pp. 190-203.

An Australian study of Higher School Certificate (HSC) results and success at Melbourne University for arts and science students. Dunn found that state school students do better at university than independent school students with the same HSC score.

Epstein, I. "An Analysis of the Chinese National Examination: The Politics of Curricular Change". Peabody Journal of Education 59 (1982), pp. 180-189.

The Chinese National Examination System was introduced in 1977. It had a marked effect on educational policy: (1) Middle school was expanded from 2 to 3 years to allow for more concentrated teaching of the exam subjects; (2) the learner comes to assume a more passive role and is expected to adapt to a preconceived learning environment (p. 186); (3) access to university is still unequal since urban facilities were better than rural; (4) pressures increased for compulsory testing at the middle school level as a prerequisite; (5) ability grouping was reinstituted at every age level; (6) certain key schools received disproportionate public funding; (7) the reputations of schools and teachers is influenced by the percentage of students who pass the exam and continue on to university; (8) teachers are forced into retraining programs for the newer curriculum; (9) failed candidates internalize a sense of failure and contemplate suicide; (10) the exams have reinforced the distance between the roles of teacher and student, teacher expertise has become more highly valued, and student dependency has been reinforced.

An historical approach, with no actual empirical results reported.

Glasman, N.S., and Biniaminov, I. "Input-output Analyses of Schools". Review of Educational Research 51 (1981), pp. 509-539.

This paper advances a useful model of school, student input and student output variables (p. 536). It may suggest some useful measures.

Glossop, J.A., and Roberts, C. "An Exploratory Study of Examination Policy Differences and Performance in Three Comprehensive Schools". Educational Review 32 (1980), pp. 67-85.

This is a U.K. study of 3 schools in a working class district, with n (of students) = 2012. Social Background Factor Score was used as a variable (individual measures comprising it are listed on page 71). Success at O Level was found to be related more to measured intelligence and social background score than to school policy, which varied for each of the three schools.

Gray, J. "A Competitive Edge: Examination Results and the Probable Limits of Secondary School Effectiveness". Educational Review 33 (1981), pp. 25-35.

Gray analyzed British data bases (e.g., Inner London Education Authority, Sheffield, and the Scottish Education Data Archive). Looking at the predictive power of intake variables, the "best" predictors were verbal reasoning scores of intakes, the social disadvantage of the schools, and the percentage of pupils from non-manual occupations; all with an r of 0.85 or greater. Controlling for differences between intake variables, differences remained that were attributed to the effects of schools themselves (p. 33).

Interesting quotation, but no reported evidence: "It is easy to forget that the examination system was originally designed for a relatively small elite of pupils and that it has grown in a somewhat topsy-turvy manner to cover a much larger proportion of pupils than was originally intended. Some teachers argue both that it has a distorting effect on the curriculum and that it fails to cater for the needs of less able pupils." (p. 27)

Haertel, E., and Calfee, R. "School Achievement: Thinking About What to Test". Journal of Educational Measurement 20 (1983), pp. 119-132.

Discussion of the validity of tests to assess instructional outcomes.

Halpin, G., and Halpin, G. "Experimental Investigation of the Effects of Study and Testing on Student Learning, Retention, and Ratings of Instruction". Journal of Educational Psychology 74 (1982), pp. 32-38.

This study showed that students who studied for and took a test achieved more and retained it longer than students who studied in order to learn rather than for a test.

Hatala, R.J. "Testing in Perspective". New Directions for Testing and Measurement 16 (1982), pp. 141-145.

He maintains that the issue is not "to test or not to test", but "Who tests, and for what reasons?". He discusses two assumptions of college admissions tests. The first is that the progression from high school to college to a profession forms a closed system. That is, admissions tests are seen as being valid for young students who have continued to progress through the "academic sequence" from high school to university, but of more limited application for older people who stepped out of the system and seek reentry. A second assumption is that verbal skills and knowledge are central to college education, the mathematics aptitude test becoming a significant predictor only for students in math-oriented programs.

Herman, J.L., and Dorr-Bremme, D.W. "Uses of Testing in the Schools: A National Profile". New Directions for Testing and Measurement 19 (1983), pp. 7-17.

This is a U.S. study of 91 school districts, conducted to see how tests were used. Principals were found to use tests for decision making and communications, although they relied more on teacher opinions and recommendations.

Teachers' attitudes to tests were described as follows: testing is a technique to motivate students to study harder; minimum competency tests (MCT) are often unfair to particular students; MCT may affect the amount of time that can be spent teaching subjects or skills the tests do not cover (thus, narrow the curriculum); teachers did not wish to be judged by students' performance on standardized tests; they did not want to be held accountable for students' scores; they had concerns about the equity of MCT for some students, and reservations about the pressure that testing exerted on teachers.

One impact of testing the authors discuss is the effect on non-tested subject areas. "Admittedly, tests alone have not caused the curriculum to narrow. Rather, the narrowing is a consequence of the importance ascribed by society at large to test scores and of an emphasis on basic skills" (p. 15). The authors query whether the sample of skills assessed by the tests represents an adequate curriculum, and whether test developers ought to be defining the curriculum.

The attitudinal test used to tap teachers' views was not included in the report. It may be available from the Centre for the Study of Evaluation at UCLA.

Lewis, D.M. "Certifying Functional Literacy: Competency Testing and Implications for Due Process and Equal Educational Opportunity". Journal of Law and Education 8 (1979), pp. 145-183.

A discussion of legal implications; test reliability, validity and instructional match (p. 159); and racial discrimination (p. 165).

Ligon, G.D. "Preparing Students for Standardized Testing". New Directions for Testing and Measurement 19, (1983), pp. 19-27.

A discussion of 8 factors that inhibit optimal performance on tests: the first four are "targets of preparation activities" (test anxiety, carelessness, confusion and poor use of time); the fifth is considered a controversial topic (unsuccessful guessing); and the last three are considered long-range issues (lack of skills, special circumstances, and handicapping conditions).

Linn, R.L. "Testing and Instruction: Links and Distinctions". Journal of Educational Measurement 20 (1983), pp. 179-190.

A discussion of four features of minumum competency tests used in classrooms. (1) Matching (the degree of match between test items and instructional objectives); (2) Feedback (the use of test results to provide feedback to students and teachers); (3) Flagging (the use of tests to flag facts or concepts that are considered important); and (4) Grading (the use of tests to determine grades). Since classroom uses of tests are stressed, this article may be of limited value to our study.

Madaus, G.F. "Testing and Funding: Measurement and Policy Issues". New Directions for Testing and Measurement 1 (1979), pp. 53-61.

Mention is made of Gallup surveys in 1976 and 1978 in which the public agreed that state or national exams should be required for high school graduation.

Madaus, G.F. "Reactions to the Pittsburgh Papers". Phi Delta Kappan 62 (1981), pp. 634-636.

A comparison of two studies of the uses made of standardized tests: the Carnegie-Mellon study in Pennsylvania and Madaus' study in Ireland. Results were similar: administrators did not use test results because they had no effect on policy. However, it is claimed that whenever test results become a key element in decisions affecting individual life chances (e.g., grade promotion), the administering agency assumes a great deal of power over the schooling process, and administrators, teachers and pupils modify their behavior and attitudes.

The testing programs mentioned were felt to lack this important dimension: that is, they "were not used as an administrative mechanism in an important policy context" (p. 635). Thus, the findings had to be interpreted with this in mind.

__________. "The Clarification Hearing: A Personal View of the Process". Educational Researcher 11 (1982), pp. 4, 6-11.

This article is about the National Clarification Hearings on MCT. Suggestions for a modified judicial evaluation model were advanced. (This was the Con view: for the Pro view, see the reference to Popham, 1982.)

__________, and Airasian, P.W. "Issues in Evaluating Student Outcomes in Competency-based Graduation Programs". Journal of Research and Development in Education 10 (1977), pp. 79-91.

Teachers and students adhere to objectives implicit in external exams rather than explicit, curricular objectives. Exams determine the instructional emphasis because they "have some import for pupils and teachers" (p. 83). The authors refer to Bloom (1969), who pointed out that "examinations which are used to make important decisions at major disjunctions in the educational system have great effects" (p. 84).

Some of the disadvantages of external exams include: the narrowing of teaching and learning; the focusing of study to the point of cramming; the mechanization of teaching and learning; the emergence of schools within schools, with pupils grouped according to whether they will likely be certified (p. 84).

In terms of the effect on the curriculum, they state that "when there is a choice between emphasizing tested or nontested objectives, it is general experience that the objectives actually tested assume primacy" (p. 85). "Most studies have found that the proportion of instructional time spent on various objectives was seldom higher than the predicted likelihood of their occurrence on the external examination."

One way that certifying exams came to control the curriculum was through their emphasis on recall of factual material (p. 86, references listed).

Regarding teacher practices, they say the principal, negative effect of external exams is that so much teaching time can be devoted to coaching or cramming for the certifying tests. There are tremendous social pressures on teachers, and "one hidden agenda in the competency-based approach is teacher accountability" (p. 88).

_________; Kellaghan, T.; Rakow, E.A.; and King, D.J. "The Sensitivity of Measures of School Effectiveness". Harvard Educational Review 49 (1979), pp. 207-230.

They compared standardized tests and public examination performance, and found that the latter, curriculum-based tests, were more sensitive to differences in school characteristics. School variables that proved to be important predictors of achievement were those reflecting the climate or activities of the school, rather than such static characteristics as size, teacher qualifications, etc. (p. 225). "What seems important in affecting achievement are the academic demands of courses, the students' concern for and commitment to academic values, the amount of time spent on study and homework, and, in general, a climate of high expectations on the part of students and their teachers."

The appendix lists the predictor variables, grouped under the headings: individuals, classrooms, family background, individual-classroom, and IQ.

_________, and McDonagh, J.T. "Minumum Competency Testing: Unexamined Assumptions and Unexplored Negative Outcomes". New Directions for Testing and Measurement 3 (1979), pp. 1-14.

Several interesting conclusions are drawn:

(1) Other states introducing MCT should look to Florida's experience and (a) not call the test a MCT; (b) not use social promotion through the grades and then use the test to deny pupils a diploma years later; (c) not introduce the tests until the measured skills are part of an instructional program; and (d) not introduce tests abruptly and without taking into account the prior educational history of the first class to be affected.

(2) When legislation weds tests to graduation, putting a test under state rather than local control results in a shift of control over schooling to the state. Two dimensions that give a test its importance as an administrative device are indicated in Figure 1 (p. 5). These are: the extent to which the local district or state controls the testing process, and the degree of impact of the test results on an individual's life chances. Different types of testing in the resulting four quadrants are then described. Provincial exams would fall primarily into quadrant I, along with most external examining. (This may be an interesting model for our study.)

(3) The history of certification exams in Europe is worthy of closer examination.

McClung, M.S. "Are Competency Testing Programs Fair? Legal?". Phi Delta Kappan 59 (1978), pp. 397-400.

Curricular validity (curricular intent) is differentiated from instructional validity (curricular reality). Also, the fairness of the tests is questioned for the following three reasons: (1) students did not receive sufficient notice prior to the phasing in of minimum competency tests as a graduation requirement; (2) the test items did not match classroom instructional objectives; and (3) students from different racial backgrounds were discriminated against.

Murphy, R.J.L. "Reliability of Marking in Eight GCE Examinations". British Journal of Educational Psychology 48 (1978), pp. 196-200.

Differences in reliability were related to: (1) subject areas; (2) number of questions in the exam; and (3) question types other than free response.

__________. "Teachers' Assessments and GCE Results Compared". Educational Research 22 (1979), pp. 54-59.

Teachers' rankings were different than the GCE exam rankings, but there was a high level of agreement between teachers' predictions of students' GCE grades and actual GCE grades.

__________. "A Further Report of Investigations into the Reliability of Marking of GCE Examinations". British Journal of Educational Psychology 52 (1982), pp. 58-63.

Marking reliability was higher than in the previous (1979) study, and higher than anticipated for essay and free response type questions.

Popham, W.J. "Melvin Belli, Beware!". Educational Researcher 11 (1982), pp. 5, 11-15.

Pro MCT at National Clarification Hearings (see Madaus, 1982, for Con argument); an appraisal of merits of MCT.

__________, and Lindheim, E. "Implications of a Landmark Ruling on Florida's Minimum Competency Test". Phi Delta Kappan 63 (1981), pp. 18-22.

Question: what is a "fair" test? Discusses content validity and the source of data required to answer the question: i.e., 1) instructional materials; and 2) classroom transactions. What is needed is a balance between the two. Comment that in the early stages of MCT, teachers will not have changed the curriculum focus to fit the testing emphasis.

Popham, W.J., and Rankin, S.C. "Minimum Competency Tests Spur Instructional Improvement". Phi Delta Kappan 62 (1981), pp. 637-639.

It is asserted that the public wants reassurance of educators' effectiveness, while educators don't want to be held accountable for pupil deficiencies that are due to external social forces.

Powell, B., and Steelman, L.C. "Testing for Sex Inequality in Standardized Admission Exams: The Case for Open Access". Integrated Education 20 (1983), pp. 86-88.

Resnick, D.P. "Testing in America: A Supportive Environment". Phi Delta Kappan 62 (1981), pp. 625-631.

An historical analysis of testing.

Resnick, L.B. "Introduction: Research to Inform a Debate". Phi Delta Kappan 62 (1981), pp. 623-624.

The introduction to a series of articles by D. Resnick, Sproull and Zubrow, Salmon-Cox, Madaus, and Popham and Rankin.

Riggs, R.O., and Lewis, W.L. "The Influence of Mandated Minimum Competency Testing on Teacher Education Curricula". Phi Delta Kappan 60 (1979), pp. 751-752.

A survey of teacher education curricula revealed that there was not as much change due to the introduction of MCT as had been anticipated.

Salmon-Cox, L. "Teachers and Standardized Achievement Tests: What's Really Happening?". Phi Delta Kappan 62 (1981), pp. 631-634.

A report of the results of a survey of 68 elementary teachers' use of standardized test data. 50% used it to confirm, 20% to guide for instructional change. When standardized test scores were lower than class performance, teachers tended to disregard test scores; when test scores were higher, teachers paid more attention to them.

Serow, R.C., and Davies, J.J. "Resources and Outcomes of Minimum Competency Testing as Measures of Equality of Educational Opportunity". American Educational Research Journal 19 (1982), pp. 529-539.

The report of a study of 1731 students in North Carolina. "Negative outcomes, mainly in the form of test failures, occurred disproportionately among blacks, and reading remediation appeared to be less effective for blacks than whites."

"From the perspective of racial equity, the major problem suggested by the present findings is that schools may be applying universalistic standards in allocation of remediation when many black pupils seem to require additional assistance." (p. 537)

"What may be needed are: earlier identification of academic deficiencies, more intensive remediation, and instruction which is directly targeted to the needs of low-achieving minority pupils."

Sharp, A., and Thomson, G. "Performance in External Examinations and Pupils' Orientations to Studying". Educational Review 36 (1984), pp. 37-51.

A U.K. study of 539 pupils in 4 schools; both quantitative and qualitative data were collected. Attitude inventory, modified from Entwhistle (1979), resulted in 5 scales: (1) Motivation and organisation; (2) Rationalising or blaming; (3) Home support; (4) Passive learning; and (5) Exam coping or strategist approach. (p. 40)

"The evidence suggests that the most successful pupils are more motivated by learning, less given to blaming and rationalisation, perceive themselves as being in receipt of more home support and better able to cope with examinations than their failing counterparts." (p.44)

Sproull, L., and Zubrow, D. "Standardized Testing from the Administrative Perspective". Phi Delta Kappan 62 (1981), pp. 628-631.

An intensive, small-scale study in Pennsylvania which reported that testing and test-related matters did not have high priority in most school systems. Also, administrators did not depend heavily upon test data for decision making.

Trusz, A.R., and Parks-Trusz, S.L. "The Social Consequences of Minimum Competence Testing". Educational Studies 12 (1981), pp. 231-241.

The authors contend that minimum competency testing "revitalizes the social role sorting function of public education in such a way as to destroy equality of educational opportunity" (p. 231). They argue that tests, by their very nature, must discriminate in favor of some groups and against others; also, the logic of test construction means that the tests must be biased in favor of a "normal" group. The inevitable effect is a tracking along class, cultural and racial lines. They discuss adjusting the difficulty of a test, and refer to 1978 statistics for three states which showed discrepancies between failure rates for black and white students (e.g., in North Carolina, 85% of blacks and 37% of whites failed). They refer briefly to a situation in which tests were adjusted in difficulty, and the ratios of black-to-white failure actually increased.

Tumin, M.M. "The Functions of Testing". New Directions for Testing and Measurement 9 (1981), pp. 21-29.

A general discussion of university admissions testing.

Tyler, R.W., et al. "Impact of Minimum Competency Testing in Florida". Today's Education 67 (1978), pp. 30-38.

A panel report. "From the testimony presented at the hearings, we conclude that the implementation has been faulty, particularly in lack of adequate communication, lack of careful consideration of all important effects of such a program, lack of planning to try to reduce or eliminate undesirable effects, and lack of decentralization to the school building level of decisions that seriously affect teachers, students, and parents" (p. 33).

"It appears as if the current class of eleventh graders who are Black and poor were sacrificed for the purpose of rapid implementation of the functional literacy segment of the Accountability Act. It is evident that there was little active concern for the appropriateness of the testing program for a large segment of the school population (the Black and the poor)" (p. 35).

"One serious potential abuse that is emerging as a result of the Florida Accountability law is the use of students' scores on the basic skills and functional literacy tests as the major criterion for evaluating a teacher's effectiveness in the classroom" (p. 36).

Weir, A.D. "The Scottish Certificate of Education: Factors Affecting Pupil Performance". Scottish Educational Studies 7 (1975), pp. 5-14.

There are possible useful suggestions here for measures of socio-economic status.

Willmott, A. "Assessment and Performance". Oxford Review of Education 4 (1978), pp. 51-64.

Defines "Assessment" and "Performance", then discusses item banking, saying it fulfills the necessary conditions for sound measurement. Refers also to the Rasch model.

B. Books, Published Reports and Unpublished Documents

Ainsworth, M.E., and Batten, E.J. The Effects of Environmental Factors on Secondary Educational Attainment in Manchester. London: Macmillan, 1974.

This is a report on the second phase of a longitudinal study of 2348 primary school children in Manchester. The intention was to identify those environmental factors associated with the students' "differential ability to extract the maximum advantage from the school system as it exists". Their study of school environment showed factors in the home background to be of major importance in educational attainment. For our study, the nine Appendices may be worth examining. These include copies of questionnaires, interview schedules, a school characteristics inventory for students, and tables of the 90 school variables and 116 individual variables, together with their source and basis for analysis.

Airasian, P.W.; Madaus, G.F.; and Pedulla, J.J., eds. Minimal Competency Testing. Englewood Cliffs, N.J.: Educatonal Technology, 1979.

Of some interest is a discussion of models under which minimal competency testing programs may be implemented (see Chapter 11, p. 183-206). Three central aspects are considered to be: specification of competencies, selection of testing procedures, and definition of standards of competency. The policy implications of the two predominant models (i.e., state administration and local district administration) are discussed in terms of seven areas: (1) state versus local control, (2) impact on curriculum, (3) remediation, (4) standards, (5) measurement issues, (6) legal issues, and (7) costs.

The editors conclude with a set of 12 recommendations for educators, parents and concerned citizens confronted with the prospect of a minimal competency program in their state or school district (p. 216-217).

Alberta Education. Revised Provincial Student Evaluation Policy. Discussion Paper. Edmonton: Alberta Education, 1983.

This position paper proposed five changes to student evaluation in Alberta: (1) compulsory comprehensive examinations for high school graduates in four discipline areas: Language Arts, Social Studies, Mathematics and Sciences; (2) increase in credit requirements in social studies for graduation; (3) discontinuation of achievement testing at the high school level; (4) transcript reporting of the comprehensive exam mark as a percentage, along with the school-awarded mark; and (5) increase in the minimum requirement for awarding a course credit from 40% to 50%.

Bloom, B.S. "Some Theoretical Issues Relating to Educational Evaluation". In Educational Evaluation: New Roles, New Means, edited by R.W. Tyler, pp. 26-50. Chicago: University of Chicago Press, 1969.

Of particular interest is his discussion of non-specified outcomes of instruction, maximal and minimal effects of evaluation, and positive and destructive efffects (see pages 38-50).

Canadian Teachers' Federation. Province Wide Student Assessment Programs. Discussion Paper. Ottawa: Canadian Teachers' Federation, 1982.

It is interesting that the date within this five-page report is February 1981, while the date on the title page is April 1982. The general concern is that teachers have no decision making power in establishing assessment practices in schools. The paper discusses the problems which arise for teachers and students, and questions the benefit of testing programs on student learning. For example, "If music and art are not subject to province-wide testing, does this mean they are not important?" and "Will schools be forced to emphasize the teaching of skills for credentials over teaching for social competence?" (p. 2) The principles and recommendations stress that teacher organizations must begin to assess the use and implications of student assessment programs.

Christie, T., and Forrest, G.M. Defining Public Examination Standards. London: Macmillan Education, 1981.

Contains a discussion of the dual functions of examinations as summaries of current attainments and as predictors of future performance. The study itself explores the "nature of the judgement that is required when examination boards are charged with the responsibility of maintaining standards". The authors discuss the importance of maintaining equilibrium between "the definition of attainment by reference to a syllabus and by reference to the performance of other candidates". Two theoretical models of grading are then considered from the point of view of their fit to models of the nature of educational achievement. A third model - limen-reference assessment - is derived, which is thought to represent current practice in public examination boards; its properties and potential development are discussed.

Council of Ontario Universities. Experimental Achievement Testing Programme: Summary Report Toronto: Council of Ontario Universities, 1979.

The objective of this testing project was to evaluate the usefulness of standardized achievement tests for assessing pre-university academic achievement and aptitude, for admission selection or for post-admission diagnostic and placement purposes. The basis of the project was that the Ontario university community felt that entering students were ill-prepared in English and Mathematics. Standardized tests in English and Mathematics were administered to entering freshmen at four Ontario universities in 1975. Because of experimental difficulties, firm conclusions could not be drawn. However, "the participating institutions did obtain benchmark data which (with some limitations) could be useful for comparison in any future administrations of these tests" (page vii).

Dumont, F.J. Alberta Grade 12 Examination Study: A Study Commisioned by the Minister's Advisory Committee on Student Achievement (MACOSA). Edmonton: Province of Alberta, 1977.

This study attempted to answer a question posed by the Alberta Legislative Assembly in 1976; namely, "What has happened to the quality of education since compulsory grade 12 examinations were dropped in 1973?" To fulfil its mandate, the study addressed three questions, the answers to which are listed below.

(1) In regard to current grade 12 student evaluation policies and practices at the school system and level across Alberta, it was found that almost all of the systems offering grade 12 had developed an implicit or explicit grade 12 student evaluation policy.

(2) Concerning the changes that had taken place in the distribution of marks in grade 12 subjects over the past five years, there were two conclusions. First, there had been an increasing trend in marks for two years after the provincial exams were dropped in 1973 and then a decreasing trend for the next two years, with an anticipated return to the average established under the former exam system. Secondly, the data indicated a difference in evaluative criteria between schools in the same system as well as among schools.

(3) To survey the public's concerns for the quality of education, 4,476 respondents were polled. The various educational groups saw a lack of common standards as being a problem since the compulsory exams were dropped. Also, the 'non-educationist' groups wanted a return of the compulsory system, and preferably a system that involved a mixture of teacher marks and departmental marks.

Dumont, F.J. Alberta Grade 12 Examination Study: Condensed Version. A MACOSA Study. Edmonton: Province of Alberta, 1977.

See above entry.

Dumont, F.J. Alberta Grade 12 Examination Study: Executive Summary. A MACOSA Study. Edmonton: Province of Alberta, 1977.

See above entry.

Dunn, S.S., ed. Public Examinations: The Changing Scene. Adelaide: Rigby, 1977.

A collection of 10 papers. The third, by Moore, on functions of examinations, is particularly thoughtful. Figure 3.1 indicates major effect areas and their constituent simpler effects. Worth perusal.

Elley, W.B., and Livingstone, I.D. External Examinations and Internal Assessments. Wellington: New Zealand Council for Educational Research, 1972.

A discussion of the advantages and disadvantages of external exams and of the accrediting system in New Zealand (p. 37-43) may be of interest.

Entwistle, N. Styles of Learning and Teaching. Chichester: John Wiley, 1981.

Contains an interesting model of factors influencing the learning process (see Figure 12.1, p. 247), and a discussion of effects of examinations on student learning (p. 261).

__________, and Wilson, J.D. Degrees of Excellence: The Academic Achievement Game. London: Hodder & Stoughton, 1977.

Etkin, B., and Leathem, B. Grade 13 Marks as a Predictor of Performance in Engineering (Parts I and II). Toronto: University of Toronto, 1978.

See paper by R. Traub (1979).

Harrison, A. Profile Reporting of Examination Results. London: Methuen Educational, 1983.

This report is concerned with how the results of public examinations in Britain could be presented in greater detail to provide more information about different kinds of achievement within a subject. That is, profile reports would show grades or marks achieved in separate components of a single examination. Several topics under discussion are reliability, methods of assessment, and uses of results.

Holmes, E.G.A. What Is and What Might Be: A Study of Education in General and Elementary in Particular. London: Constable, 1911.

Holmes recognized the dangers of evaluating a program according to the most readily measured aspects, while the objectives of the program are not reflected by the measuring instruments. Interesting quote contained on page 128.

Kellaghan, T.; Madaus, G.F.; and Airasian, P.W. The Effects of Standardized Testing. Boston: Kluwer-Nijhoff, 1982.

This study examined the effects of standardized norm-referenced tests. The authors chose to conduct their investigation in Ireland, a country in which very little use had been made of standardized testing in the past. An experimental design was used in which some schools would test and receive test information, while others would not. The main features of an American testing program were simulated. In general, it was concluded that standardized testing tended to be used to support, rather than to disrupt, existing school and teacher practices. This study is also discussed by the above authors in journal articles cited above (see Section A).

Livingstone, D.W., and Hart, D.J. Public Attitudes Toward Education in Ontario: 1979. Toronto: OISE Press, 1980.

Of interest to our project is the public response to the introduction of examinations. The first dealt with control over curriculum development, and the second, with testing the academic progress of students in the higher grades (p. 20).

Montgomery, R. A New Examination of Examinations. London: Routledge and Kegan Paul, 1978.

This book is an historical/philosophical discussion. Chapter 2, on the functions of examinations, may be of limited interest for our study. Once again, the distinction is made between qualifying exams and competitive exams.

Moore, W.E. **"Some Functions of Examinations".** **In** **Public Examinations: The Changing Scene,** **edited by S.S. Dunn, pp. 51-75. Adelaide: Rigby, 1973.**

An insightful article on the effects of public examinations, with main effect areas defined as examination roles, goals, and the curriculum. Figure 3.1 (p. 54) indicates major effect areas and their constituent simpler effects.

Queensland Department of Education. Public Examinations for Queensland Secondary School Students. Brisbane: Queensland Department of Education, 1970.

The pro's and con's may be of interest (p. 54-58). Also, they refer to a dated study of 400 freshman entering the University of Queensland in 1955 in which matriculation scores and IQ results were found to be not significantly related to university performance.

Ratsoy, E.W. Public Reactions to the Proposed Provincial Student Evaluation Policy. Edmonton: Alberta Education, 1983.

This report presents an analysis of public responses to a request for submissions by Alberta Education upon release of a discussion paper on a proposed revision to the provincial student evaluation policy. (See comments on this paper under entry for Alberta Education.) At the time (March 1983), Grade 12 comprehensive examinations were optional exams in the Alberta system and the discussion paper recommended that they be made compulsory.

Ratsoy categorized five public groups who wrote submissions: parents and other lay groups, in-school groups, school jurisdictions, Alberta associations, and post-secondary groups. Their reactions to six proposed modifications in the examination system were summarized as well as twenty related concerns. These are of interest to us in terms of public reactions to educational standards.

Reid, J.E. "Inflation of Standards: Fact or Fiction?". A paper presented at the Annual Meeting of the Canadian Educational Researchers Association, London, Ontario, 1978.

This is a very useful article. (1) Shows grade inflation between 1966 and 1976 (departmentals were discontinued in 1973); (2) Teachers better "qualified" in 1976 than earlier; (3) Student competence, as measured by common exams, was steady over time (1975-1977); (4) Public perception is that standards are declining; (5) Survey respondents in an "educator" category were against return to exams becauuse: a) evaluation should be based on a year's work, b) multiple-choice format is too restrictive, and c) non-exam subjects would be treated as second-class; (6) If there is a return to exams: a) curriculum objectives and evaluation procedures should be specified early in the year and communicated to the student, b) evaluation procedures should include more than multiple choice tests, and c) the system should involve a combination of teacher and exam marks.

Stewart, C., and Rhodes, H.C. eds. "Achievement Test Scores Show a Decline in U.S.: A Summary of Journal Articles and an Annotated Bibliography". Alberta Education, 1976. (unpublished mimeograph)

Strenio, A.J. The Testing Trap. New York: Rawson, Wade, 1981.

Contains an interesting, conversational chapter on the impact of test abuse.

Traub, R.E. "Unsupported and Iniquitous: A Proposal by Bernard Etkin and Brian Leathem". Commentary prepared for Research and Evaluation Branch, Ontario Ministry of Education, 1979.

This is a commentary on the two-phase report by Etkin & Leathem (see earlier citation).

__________, and McLean, L.D. "A Rosy View -- University Admission Officers' Preferences and Expectations for Provincial Examinations". The Ontario Institute for Studies in Education, 1984. (unpublished report)

Traub and McLean discuss different pressures on the Ontario government for a return to some kind of province-wide testing.

__________; Wolfe, R.; Wolfe, C.; Evans, P.; and Russell, H. Secondary-Postsecondary Interface Project II: Nature of Students. Volumes I and II. Toronto: Ministry of Education and the Ministry of Colleges and Universities, Ontario, 1977.

The purpose of the set of three Interface projects was to examine the interface between schooling at the secondary level and postsecondary education. This project focused on the nature of students and their achievement; the other two dealt with perceptions and the curriculum. The most interesting result concerned the predictability of the first year mark average in university. Accounting for differences in marking standards among universities and among program areas, the correlation coefficient between Grade 13 mark average and the first year mark average was 0.64. A second comparison of teacher marks and achievement test performance indicated some evidence of mark variation.

Tyler, R.W., ed. Educational Evaluation: New Roles, New Means. The Sixty-eighth Yearbook of the National Society for the Study of Education. Part II. Chicago: Unversity of Chicago Press, 1969.

This is a general collection of articles. The most relevant might be Chapter III by B.S. Bloom, "Some Theoretical Issues Relating to Educational Evaluation" and, in particular, pages 38-50 in which he discusses non-specified outcomes of instruction, maximal and minimal effects of evaluation, and positive and destructive effects.

Webber, C.F. "School Board Member Perceptions of the Utility and Importance of Student Evaluation Information in Alberta". Planning Services, Alberta Education, 1984. (unpublished report)

This survey was conducted on 196 school trustees to determine how well the Alberta provincial test data were being communicated and how useful the data were for decision making. Some conclusions were: (1) the data were perceived as being most useful for maintaining educational standards, (2) trustees wanted to use test results to compare schools and teachers, as well as students, and (3) the initial impact of the exams seemed to be minimal.

Wigdor, A.K., and Garner, W.R., eds. Ability Testing: Uses, Consequences, and Controversies. Washington, D.C.: National Academic Press, 1982.

A report by the Committee on Ability Testing intended to describe the theory and practice of testing and to help decision-makers make better-informed judgments about tests and test use. Part I contains seven chapters, including the topics: methods, historical and legal context, employment testing, ability testing in elementary and secondary schools, and admissions testing in higher education.

Part II, the documentation section, is a set of 11 papers organized under three headings: employment testing, educational testing, and psychometric issues. Gardner's paper on use and misuse of tests, and Cole's paper on the implications of coaching might provide some insights for our study.

C. Newspaper Articles

"Blacks Shy Away from O Levels". The Times Educational Supplement 3157, December 5, 1975, p. 13.

In a study of 668 students, each white had an average of 2.8 passes, each black 1.9. Reasons given by schools for lack of achievement (including language and linguistic problems, and lack of discipline) were not supported in the research findings.

Doe, B. "Quarter of Candidates Get the Wrong Marks". The Times Educational Supplement 3155, November 21, 1975, p. 3.

Discussion of a report, "The Reliability of Examinations". Where grades are assigned on a 5-point scale, unreliability of measurement means that 25% of grades may be under- or over-estimated. The report looked for internal consistency (reliability) by treating different parts of the exams as if they were separate measures of the same thing. The typical reliability coefficient for the exams was 0.88.

________. "Choice of Questions Confuses O Level Marking". The Times Educational Supplement, 3158, December 12, 1975, p. 5.

Discussion of a report, "O Level Examined: The Effect of Question Choice." Report alleges 2 out of 5 might get different grades if questions were fairer. There is a need to equalize questions, otherwise ability to choose questions rather than ability to answer is being assessed. Also, it was noted that in marking, the proportion of scale used was different for different subjects (Math:75% of scale, English: 45%). Concluded that profiles were better than adding marks from several papers measuring different aspects of the subject.

________. "Exam Board Joins the Don't Knows in Debate over Shifting Standards". The Times Educational Supplement 3242, July 22, 1977, p. 3.

The topic in this article is the problem of defining exactly what "examination standards" are. Wider range of curricular choice may be the cause of reported inconsistencies of standards. If standards are difficult to define, then a "fall in standards" must be even more difficult to pinpoint. It might refer to pupils being less skilled or knowledgeable, teachers being less hardworking or competent, examiners more erratic, or today's curriculum less worthwhile. "It is not for the examining boards to decide whether consistency in public exams should take precendence over freedom in the curriculum, but it is an issue to be faced in the discussions taking place about core curricula and educational standards".

________. "Grading System May Have Caused Inaccurate CSE Results". The Times Educational Supplement 3326, March 7, 1980, p. 8.

Method recommended to moderate teacher assessment may have caused one third of papers to be misgraded. To check on exams set and marked by teachers, an external examiner marked 20 papers from each teacher and modified the teacher grade if there was a half grade difference between the examiner's and the teacher's average. See Journal article by Boreham (1979) also.

Fields, C.M. "Measuring the Impact of Standardized Tests". The Chronicle of Higher Education 17, November 27, 1978, p. 13.

Report on public hearings held by the National Academy of Sciences' Committee on Ability Testing. One issue was the over-reliance on examinations for admissions, placement and hiring decisions; critics felt that "standardized tests contain racial, class, sex, or geographic biases". (A report was scheduled to be issued January 1980)

"Heads Angered by Use of Figures for Political Ends". The Times Educational Supplement 3299, September 22, 1978, p. 5.

Reaction to publication of a table of A level results of individual comprehensive schools in Manchester. Report looked at possible reasons for a long-term decline in A level pass rates, and suggested that social and demographic changes in the inner city were altering the ability range of students. One critic, who said the figures were misleading commented: "There are fewer bright children in deprived working-class areas than in affluent middle-class ones. The free parental choice system now operating in Manchester also means that ambitious parents tend to send their children to (other) schools." (Sending children to other schools is a possible effect to be noted.)

Kirkaldy, J. "Employers Unhappy with State Test". The Times Educational Supplement 3300, September 29, 1978, p. 18.
Australian report. Employers there want a return to external exams for school certificate. Teacher assessments are too difficult to interpret.

Makins, V. "Why Cream of Sixth Goes Sour". The Times Educational Supplement 3243, July 29, 1977, p. 3.

Report on "Degrees of Excellence" by Entwistle and Wilson. 'A' level exam results are poor predictors of success at university, more especially for arts subjects and social sciences, and less for mathematics. One cause of students' difficulties may be the mismatch between sixth form and university teaching styles. "In the sixth form ... the pupil's working habits are controlled by the teacher, there are generally many small pieces of work done on a regular basis ... Teachers and pupils work together at overcoming a 'common enemy' - the external examiners." (Will exams make closer allies of teachers and students in courses with exams?)

"At university, the student ... may find large first-year classes, ill-defined course objectives, infrequent assessment of progress and many lecturers who do not primarily consider themselves as teachers. In addition, students and staff find themselves on opposite sides of the assessment 'war'."

_________. "Exams Cause of Narrow Teaching". The Times Educational Supplement 3491, May 27, 1983, p. 12.

This article is based on HMI reports (available from the Department of Education and Science) on several comprehensives and one prep school in the London area. Based on the above reports, the writer draws these conclusions: (1) Exams are causing narrow and didactic teaching; (2) Oral and aural language competence is on the decline, and the range of writing gets narrower; and (3) The needs of less able students are ignored. (However, statistics are not included here.)

Morgan, D.I. "Fresh Insights - After Which Nothing Will be Quite the Same". The Times Educational Supplement 3282, May 26, 1978, p. 15.

Regarding the proposal for N and F (Normal and Further levels) as a two-tier examination to replace the A (Advanced level) examinations in Great Britain.

Morris, M. "More Realistic for the Majority". The Times Educational Supplement 3279, May 5, 1978, p. 14.

About N and F examinations. (See previous item.)

Mortimore, P. "Why Not Marry 16-plus and Graded Tests?". The Times Educational Supplement 3465, November 26, 1982, p. 2.

The proposed examination is only for 60% of the ability range in each subject, so students will have to be divided according to ability, "despite the fact that our views on ability have broadened and our understanding of the influence of both individual motivation and school effectiveness have developed". (In terms of our study, we might want to consider the effect of exams on streaming according to ability.)

A second concern regards optional papers, and the problem of comparing a good performance on an easy paper with poor performance on a hard paper.

Passmore, B. "Welsh Pupils More Reluctant to Stay on for CSE Exams". The Times Educational Supplement 3436, May 7, 1982, p. 10.

Reports on "Public Examinations in Wales: Attainment at 16 Plus." Fewer Welsh take the CSE exam than English, and have slightly worse results. There is significant variation among the eight school districts, and among schools within a single district. The lowest attainment was from a school with the lowest percentage of professional and non-manual workers, but socio-economic factors seemed to be less important than a rural-urban difference. "The consistently good results achieved in public examinations in the rural counties may reflect the more widespread retention of traditional values and motivation in the pursuit of academic qualifications which have been eroded in some other parts of Wales, notably the industrial valleys."

However, exam results may not tell the whole story. "Some good schools have deliberately avoided examination goals in devising courses for their less able pupils in years four and five which are educationally successful, they argue."

________. "Public Exams - Main Cause of Welsh Under-Achievement". The Times Educational Supplement 3485, April 15, 1983, p. 14.

This is from the report of a conference. Exams were felt to distort the curriculum; also exam groups were said to get more than their fair share of resources. In addition, sorting pupils into exam and non-exam groups was bound to lead to absenteeism, disruptive behaviour

and lack of motivation. Graded tests, pupil profiles, and mixed ability teaching were recommended.

________. "Selective Schools Chalk up More O'level Successes". *The Times Educational Supplement* 3496, July 1, 1983, p. 3.

Report from "Standards in English Schools - An Analysis of the Examination Results of Secondary Schools in England for 1981". Pupils in selective schools pass on average a third more O levels than pupils in a fully comprehensive system, this finding based on a survey of 350,000 pupils in England and Wales. Results were adjusted to allow for social class differences.

Purvis, B. "Unfair Exam Scales Put under Review". *The Times Educational Supplement* 3310, November 16, 1979, p. 21.

Australian Higher School Certificate (HSC) marking scale is criticised for penalizing students taking certain subjects and favouring those taking subjects with a higher rating under the system. "It has been described as a vain attempt to equalize things that are not equal: the quality of teachers, pupils' ability and the subject-matter of diverse examination papers such as physics and English."

Also, the value of the HSC as a matriculation exam is in dispute. "In Victoria, the reliance on a pupil's HSC score as the sole determinant for university entry has been blamed for the high rate of first-year drop-outs."

Sayer, J. "Why Profiles are More Attractive". *The Times Educational Supplement* 3442, June 18, 1982, p. 4.

Asserts that the curriculum is dominated by the need for examination passes. "Evaluation and monitoring of the whole school process offers a richer solution than ranking the performance of individual pupils when it is too late to do anything about it."

"The Examination Tail Will Wag". *The Times Educational Supplement* 3435, April 30, 1982, p. 2.

Which comes first - examinations or the curriculum? Concern is expressed over the fact that the two are intertwined; whereas the curriculum should be paramount, the examination council is actually in control.

"Union Condemns Cramming as Unprofessional". *The Times Educational Supplement* 3179, May 7, 1976, p. 20.

A report from Ireland. With an increase in university entrance requirements, after-hours cramming schools are making a profit. The teachers' union sees 'cramming' as a breach of professional ethics. Many highly qualified secondary teachers who also hold full-time teaching positions in ordinary schools are putting in 5 or 6 extra hours a day for a profit. The union is worried about the effect of the extra workload, and the implications for the normal teaching of teachers involved in the cramming schools.

Warnock, M. "The Underlying Question: Division of Labour". *The Times Educational Supplement* 3279, May 5, 1978, p. 14.

About N and F examinations in Great Britain.

"Working-class Pupils Still Penalized". *The Times Educational Supplement* 3144, September 5, 1975, p. 12.

This refers to the journal article by Weir. A study of 2500 students in 70 schools in Scotland showed that capable students, many from a working-class background, are leaving school without attempting the O level exams and that their potential is not being realized ('potential' having been measured by a number of ability tests, occupational interest, and personality tests). Also, a developed Math test was a better predictor of O level success than the standard Verbal Reasoning Quotient, perhaps because it was less contaminated by social class factors.

D. Documents from the Educational Resources Information Center (ERIC)

Bauer, R. "Analysis of the Ohio Occupational Achievement Tests". Tau Associates, Fairmont, West Virginia, 1981. (ERIC Document No. ED 221 684)

A study of the relationship between test performance and post-high school experiences.

Beck, M.D., and Stetz, F.P. "Teachers' Opinions of Standardized Test Use and Usefulness". Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, 1979. (ERIC Document No. ED 177 202)

A national sample of 3,300 teachers who had recently administered the Metropolitan Achievement Tests responded to a questionnaire regarding test issues. Teachers rated their opinions on the *amount of such testing in their schools*, *personal use* of test results, possible test score *applications*, and other test-related *policies*.

Burry, J. *et. al*. "Teaching and Testing: Allies or Adversaries". Papers presented at the Annual Meeting of the National Council on Measurement in Education, Los Angeles, 1981. (ERIC Document No. ED 218 337)

This contains three papers from the Test Use Project of the Center for the Study of Evaluation. The third paper may be useful: "The Design of Testing Programs with Multiple and Complementary Uses". Here, interview data are discussed as examinations of educators' views about uses of assessment in external accountability and instructional decision-making.

Coffman, W.E. "Those Achievement Tests - How Useful?". Iowa University, Institute for School Executives, 1980. (ERIC Document No. ED 209 762)

Standardized achievement tests can be misused as indicators of a school's quality or effectiveness relative to other schools when variation among schools is ignored (eg., with respect to such factors as student abilities, family support of education, and student mobility); exam misuse may be an effect worthy of consideration.

David, J.L. "Local Uses of Title I Evaluations". July 1978. (ERIC Document No. ED 187 727)

This is a report of a survey of administrators, teachers and parents in fifteen school districts. Standardized achievement tests, used as the basis of evaluation, were considered inadequate for judging programs and biased when compared to skill-specific tests, observation, self-concept or attitudinal measures. Respondents disliked evaluation and ignored negative results if they believed in a program.

District of Columbia Public Schools. "Achievement in Mathematics, 1977-78, Grades One-Six. Mini-Report 79-4". Washington, D.C., 1979. (ERIC Document No. ED 182 192)

This study looked at the relationship between performance on the Prescriptive Mathematics Test (PMT) and class size, attendance, percent of males in the class, number of different test levels used in a class, and student turnover rate in the class. These factors might be considered as possible effects of examinations.

Fetler, M. "Use of Evaluation Data and School Achievement". March 1982. (ERIC Document No. ED 218 343)

Survey of elementary school principals in California who used third grade achievement data reported by the California Assessment Program (CAP). More common uses were looking for trends, examination of the curriculum, revision of existing programs, development of corrective instructional strategies, identification of new problem areas, and public communication. Two types of uses (monitoring achievement and curriculum review) had strong positive correlations with achievement, while a third type of use (textbook review) had a weak correlation with achievement.

Fillos, R.M., and Magoon, A.J. "Evaluation Acceptance in Elementary School Teachers: A Construct Validation and Description of the Meaning of Standardized Achievement Testing". Paper presented at AERA, Toronto, 1978. (ERIC Document No. ED 161 942)

Fifty third-grade teachers were surveyed about the meaning of standardized achievement testing. Subjects were from schools performing above or below prediction on the Delaware Educational Assessment Program Longitudinal Study. Multimethod-multitrait procedures showed the construct, *evaluation acceptance*, as valid, consistent and distinct from trait anxiety.

Gomez, A.S. "New Mexico Standardized Testing Program Report, 1980-81". New Mexico State Dept. of Testing, Santa Fe, 1982. (ERIC Document No. ED 227 125)

A statewide testing program at grades 5, 8 and 11 used the Comprehensive Tests of Basic Skills to measure achievement. Longitudinal comparisons were made by years and grade level, indicating improved performance for fifth and eighth grades compared to 1979 and 1980 data.

Harnischfeger, A., and Wiley, D.E. "The Decline of Achievement Test Scores: Evidence, Causes and Consequences". ERIC Clearinghouse on Tests, Measurement and Evaluation, Princeton, N.J., 1977. (ERIC Document No. ED 141 412)

Results of an investigation on the decline of scores. First, data from ten test batteries were reviewed to determine whether scores had declined since the sixties (ie., the *evidence*). Second, changes in tests and test takers were examined as potential *causes* of the decline. Third, societal and curriculum changes were examined as possible *causes*. The authors concluded that no single cause of declining scores would be identified, and that the assessment of causes was hampered by complex school and social factors.

Helsley, T. *et al*. "South Carolina Statewide Testing Program, 1980-81. Summary Report". Office of Research Report Series, Volume 1 (58), 1981. (ERIC Document No. ED 226 063)

The Comprehensive Tests of Basic Skills were used in the statewide program on 135,000 students. Longitudinal comparisons examined the same students across years from 1976-77 to 1980-81. However, changes in the baseline grade emphasis in 1980 from Grades 4, 7, and 10 to Grades 3, 6, and 11 prevent longitudinal comparisons in subject areas.

Massachusetts State Deptartment of Education. "Interpreting and Using Commercial Achievement Test Results. Basic Skills Improvement Policy: Supplement to Implementation Guide I". Boston, 1982. (ERIC Document No. ED 221 589)

This is intended to help schools get the maximum amount of useful information from test results: to help them in areas of program monitoring, grouping, planning instructional activities, and reporting results to parents and the community. (Will such supplements be a consequence of the introduction of tests?)

Popham, W.J. "Classroom Implications of Criterion-Referenced Tests: Curriculum--Instruction--Evaluation". Instructional Objectives Exchange, Los Angeles, 1976. (ERIC Document No. ED 171 762)

Discusses the inadequacies of norm-referenced tests as a basis for evaluation: (1) mismatches between local curriculum and what is measured by standardized tests; (2) anxiety and frustration caused by imprecise knowledge of needed skills improvement; and (3) tendency to omit test items which measure the knowledge teachers consider most important, in order to obtain a large degree of response variance. (These are effects of examinations that should be considered.)

Roeber, E.D. "Teaching Local Educators to Use and Report State Assessment Results". Michigan, 1980. (ERIC Document No. ED 211 570)

A set of procedures designed to help school district and building staff use and report assessment results at district, school and classroom levels. (Will anyone learn to use the provincial exam test results?)

Sherman, S.W., and Robinson, N.M., eds. "Ability Testing of Handicapped People:

Dilemma for Government, Science, and the Public". National Academy of Sciences, Washington, D.C., 1982. (ERIC Document No. ED 221 560)

A report from The Panel on Testing of Handicapped People, which examined testing and selection practices in schools and the workplace in order to describe the extent to which testing is a barrier to the full participation of handicapped people in society. They recommended necessary research into studies of test validity, validation procedures, test modifications, and investigation into the role of test scores in decision making. (What effect will provincial tests have on handicapped people?)

Yeh, J.P. "Test Use in Schools: Studies in Measurement and Methodology, Work Unit 4". Center for the Study of Evaluation, Los Angeles, 1978. (ERIC Document No. ED 214 951)

Teachers' knowledge of and attitudes toward testing were investigated. Test taking skills, test quality, and student motivation were viewed as more important factors in test scores than quality of instruction or student ability. (What effect will provincial testing have on teachers' attitudes and classroom practices?)

APPENDICES FOR PART THREE

A Review of Literature on Examinations and Assessments

Appendix B

A Selectively Annotated Bibliography on the Impact of Assessments

THE IMPACT OF ASSESSMENTS: AN ANNOTATED BIBLIOGRAPHY

A. Journal Articles

B. Books, Published Reports and Unpublished Papers

A. Journal Articles

Anderson, R.E.; Welch, W.W.; and Harris, L.J. "Methodological Considerations in the Development of Indicators of Achievement in Data from the National Assessment". Journal of Educational Measurement 19 (1982), pp. 113-124.

This is a report of a study designed to examine the utility of mathematics data from the American National Assessment of Educational Progress (NAEP) for identifying and developing indicators of mathematics achievement. This type of use of assessment results has increased in the United States since the NAEP data management was reorganized to facilitate a wide variety of secondary analyses, with NAEP data now being disseminated as a series of public-use data tapes. The authors wanted to study the implications of this NAEP data reorganization for the data analyst.

From the analysis of mathematics achievement of 17 year olds, it was concluded that the NAEP did have the potential for developing indicators of achievement in mathematics, but that interpretations of findings would have to be made cautiously because the NAEP item subsets did not consistently meet conventional psychometric criteria. The authors suggested that data analysts pay less attention to standards of item discrimination and construct validity, and concentrate on "standards of face validity, content validity, internal consistency and the application of rigorous data analysis techniques" (p. 123).

Barnes, R.E.; Moriarty, K.; and Murphy, J. "Reporting Testing Results: The Missing Key in Most Testing Programs". National Association of Secondary School Principals' Bulletin 66 (1982), pp. 14-20.

The authors suggest that achievement testing can be a dangerous activity, one reason being that results are often misinterpreted or exaggerated by various groups. This is especially important when the public perceives testing as "reflective of the value, quality, or respective effectiveness of a school or school system" (p. 14). In reporting test results, they discuss the need to identify the recipients of the testing information, and to consider the type of information each group of recipients will need. The individual needs of seven subpopulations receiving testing information of varying complexity and specificity are then outlined.

Bleecher, H. "The Authoritativeness of Michigan's Educational Accountability Program". Journal of Educational Review 69 (1975), pp. 135-141.

This is a report of a study of teachers' attitudes towards Michigan's state assessment plan, which features learning assessments at the fourth and seventh grade levels. A sample of teachers most affected by this plan (kindergarten through grade eight level) was surveyed by questionnaire. The results indicated that teachers believed the assessment to be inconsistent with the purposes of the school, and did not feel mentally and physically able to comply with the assessment. The author observed that the attitudes of teachers towards the assessment program "appear to be polarizing negatively" (p. 141), and he predicted that this group of teachers would likely withdraw their cooperation from the asssessment.

Bloom, B. "Toward a Theory of Testing Which Includes Measurement-Evaluation-Assessment". In The Evaluation of Instruction edited by M.C. Wittrock and D.W. Wiley, pp. 25-50. New York: Holt, Rinehart and Winston, 1970.

Bloom perceived testing to consist of three aspects, measurement, evaluation and assessment. He discussed each aspect in turn, and presented the view that there ought to be a synthesis of the three. Bloom perceived assessments as "attempts to assess the characteristics of individuals in relation to a particular environment, task, or criterion situation" (p. 30). Bloom stressed the need for assessments to be equally concerned with the individual as with the individual's environment.

Bock, R.D.; Mislevy, R.; and Woodson, C. "The Next Stage in Educational Assessment". Educational Researcher 11 (1982), pp. 4-11.

This paper begins with a brief review of four influences on the early development of educational assessment; these are the accountability movement in education, proliferation of survey sampling studies, development of matrix and multiple-matrix sampling, and the creation of the National Assessment (NAEP). The authors then posit that the next stage of growth in assessment will be in the area of reporting assessment results, through the development of better methods of collecting and reporting data. One such method discussed is the application of item response theory.

Datta, L-E. "Communicating Evaluation Results for Policy Decision Making". In Educational Evaluation Methodology: The State of the Art edited by R.A. Berk, pp.124-145. Baltimore, Md.: John Hopkins Unviversity Press, 1981.

This general paper on the use of evaluation results is based on the premise that conducting an evaluation implies that information is to be communicated for the purpose of making decisions. Techniques and issues in communicating evaluation findings to decision makers are then reviewed. The author argues that effective communication has to be planned at the beginning of an evaluation study as an integral part of the planning and execution.

Forbes, R.H. "NAEP: One 'Tool' to Improve Instruction". Educational Leadership 34, (1977), pp. 276-281.

Forbes maintained that the major impact of the NAEP would come only after several learning areas had been reassessed. However, in the short-term, several uses had already been recognized. This report briefly describes several cases in which the NAEP information had been used to evaluate and improve instructional programs; for example, by using modified NAEP instruments and methods to assess students' skills at the local level; by using NAEP data as rationale to obtain funding for innovative instructional aids; and by using comparative NAEP survey data (state results with both regional and national findings) to pinpoint problem areas in state curricula.

Geisert, G. "National Assessment: A Model for State and Local Competency Mandates?". Compact 13 (1979), pp. 21-23, 29.

This paper is a request by the author for states to consider the NAEP assessment model as a viable alternative to state-wide minimum competency testing. The author's opinion is that minimum competency testing is harmful to most students, and beneficial to the few for whom the testing was intended. His rationale for selecting the NAEP model over minimum competency testing includes the following points:

- After ten years of data collection, the NAEP has "become a yardstick against which to measure American education" (p. 22).
- The NAEP avoids engaging in simplistic types of educational research.
- The NAEP covers the breadth of education (including academic disciplines, individual skills and public concerns), surveys the population broadly by age (9, 13, 17 & 26-35) as well as by other demographic characteristics.
- The NAEP findings are descriptive of educational attainments by American students, assuming "no specified causal relationship between achievement/performance results and" either the academic institutions or other societal forces (p. 22).
- A very general discussion on potential uses of NAEP materials is also included.

Goldstein, H. "Measuring Changes in Educational Attainment Over Time: Problems and Possibilities". Journal of Educational Measurement 20 (1983), pp. 369-378.

Goldstein reports that, despite the fact that the APU and the NAEP gave a high priority to making inferences about trends over time, there appears to have been little attempt to define the meaning of "trends over time" or to discuss the associated measurement problems.

Using the British Assessment of Performance Unit (APU) to illustrate them, Goldstein outlines the most commonly used methods of measuring absolute change in achievement over time, and presents alternative formulations. He suggests that measuring relative changes over time, using standardized differences or longitudinal analyses would be feasible and substantively interesting.

Hextall, I. "Rendering Accounts: A Critical Analysis of the APU". In Selection, Certification and Control edited by P. Broadfoot, pp. 245-262. London: The Falmer Press, 1984.

This paper is intended to set the APU in a broad historical, social and political perspective. Hextall expresses a concern that the principles, criteria and procedures of assessment ought to be be more open to public debate, and, in fact, be decided together with those members of the public directly affected by the establishment of an assessment program. To illustrate this need, he discusses the implications and effects of certain APU decisions, for instance, deciding on the purpose of the assessment and the features of the sampling plan used.

Hiebert, J. "Units of Measure: Results and Implications from National Assessment". Arithmetic Teacher 28 (1981), pp. 38-43.

This paper is based on NAEP results about elementary students' conceptions of the unit of measure. Student responses to several exercises on the NAEP mathematics assessment are described, and then interpreted to draw conclusions about students' lack of understanding of the concepts. The author recommends that teachers administer the NAEP exercises reproduced in the article in order to confirm their own students' degree of understanding and need for clarification on the topic.

Husen, T. "An International Research Venture in Retrospect: The IEA Surveys". Comparative Education Review 23 (1979), pp. 371-385.

This is a record of personal experiences gained from the International Association for the Evaluation of Educational Achievement (IEA) studies conducted over a twenty-year period, and presented for the benefit of future large-scale surveys of this kind. An insightful and useful article, it has excellent sections dealing with the background and history of the IEA, the administrative and economic problems in setting up an international organization, and the development of the theoretical framework within which the early studies were designed. Although Husen attributed some of the weaknesses of these studies to the fact that it was pioneering, state-of-the-art research at the time, he suggests that many problems arose from the fact that the IEA researchers were restricted by the prevailing research paradigms: "the use of an input-output model, extensiveness of scope, and emphasis on quantitative methods and statistical techniques with no reliance at all on qualitative observations and anthropological methods" (p. 384).

Impara, J.C. "Valid and Invalid Uses of Statewide Assessment". Educational Technology 18 (1978), pp. 5-9.

Changes in statewide assessment concepts in the U.S. are attributed to the accountability movement of the 1960's and the inception of the NAEP. Impara gives two reasons for invalid uses of assessment: First, the "users" of assessment who are decision makers or legislators have not been educated about the limitations of test use in large-scale assessment; second, criteria for the validity of assessment use are not considered in developing an assessment program. Three criteria are suggested to determine the validity of assessment: (1) the effectiveness of the program is measured by examining whether the program goals were met; (2) the use of assessment must be well grounded and justifiable from a conceptual or theoretical perspective; and (3) the tools, such as tests and other data collection procedures must meet technical criteria for quality (p. 5).

Impara discusses several examples of invalid uses of statewide assessment, for example, making comparisons between schools or districts, and using state assessment data to allocate state school support on a district-by-district basis.

Johnson, G.H. "Making the Data Work". Compact 6 (1972), pp. 29-30.

This article can be considered more historical than explanatory, since it was written shortly after the NAEP began. Johnson considered that the eventual success of the NAEP would be measured in terms of its impact on educational practice, content and decision making; however, he felt that it was appropriate to ask whether anyone was using the NAEP. Of interest is the fact that he perceived the importance of identifying different aspects of assessment, and the fact that each aspect might have different potential uses. Two types of NAEP "products" capable of being used were categorized. The first type consisted of the model, technology and materials for conducting assessments at state and local levels; and the second type consisted of data on achievement in each assessed subject area. "The potential for use or application, and the audiences (or users) involved, may be quite different for these two kinds of NAEP products" (p. 29).

Kearney, C.P. "Uses and Abuses of Assessment and Evaluation Data by Policymakers". Educational Measurement: Issues and Practice 2 (1983), pp. 9-12, 17.

Kearney discusses two purposes of large-scale assessments, (a) public reporting at the local, state or national level, and (b) identifying needs and allocating resources. In terms of public reporting, he argues that policymakers ought to use test results to indicate to the general public what is going on in the schools. In terms of identifying needs and allocating resources, he argues that policymakers ought to be provided with the kind of test results that identify relative strengths and weaknesses at different levels in the system. He provides several examples of uses and abuses of data by policymakers with respect to these two purposes. In conclusion, Kearney calls for the establishment of a national clearinghouse of programs and practices for reporting and using test results. He claims that the establishment of large-scale achievement testing programs in almost every state during the past decade has resulted in massive amounts of information that are often not well used.

Lapointe, A.E., and Koffler, S.L. "Your Standards or Mine: The Case for the National Assessment of Educational Progress". Educational Researcher 11 (1982), pp. 4-9.

This is a follow-up to a report that evaluated the NAEP's performance and value, and suggested how the NAEP could be made more effective. (One of the co-authors of this article also participated in the NAEP evaluation). Here, the authors supply the rationale for the suggestions made by the evaluators. Of particular interest is a discussion on the NAEP's potential contribution to the "standards" debate. The authors perceive the role of the NAEP in the search for a working definition of the term "standards" as one of measuring student achievement and reporting it publicly.

Lewis, A.C. "Washington Report: Comparable Data for State-by-State Comparisons Could Become a Reality". Phi Delta Kappan 65 (1984), pp. 659-660.

Lewis discusses the problem of finding data for comparative purposes. She reports that American legislators at different levels of government have been attempting to make comparisons between states using such indicators as Scholastic Aptitude Test (SAT) scores; these were not intended to provide data on group achievement, but rather to measure individual aptitude.

Lewis suggests that a potential answer to the problem of comparative data may have been provided by a recent change in the NAEP management, from the Education Commission of the States (ECS), to the Educational Testing Service (ETS). ETS has claimed to be interested in innovation and in fostering more uses for NAEP data. In the original design of the NAEP, the sample was too small in any given state to provide state-by-state comparisons although individual states could piggyback on NAEP surveys for more extensive surveying of their own students.

Lindquist, M.M.; Carpenter, T.P.; Silver, E.A.; and Matthews, W. "The Third National Mathematics Assessment: Results and Implications for Elementary and Middle Schools". Arithmetic Teacher 31 (1983), pp. 14-19.

The results of the second and third NAEP assessments of mathematics for elementary and middle schools are compared. A major finding is that many significant gains were due to improved performance on routine exercises, and not on exercises assessing deep understanding or applications of mathematics. Specifically, Lindquist et al. elaborate on this finding by discussing whole numbers, fractions and decimals, and other basic concepts and skills. NAEP exercises (as well as national percentiles) from which performance on these content areas was interpreted are included in the article.

Madaus, G.F. "Reactions to the Pittsburgh Papers". Phi Delta Kappan 62 (1981), pp. 634-636.

A comparison of two studies of the uses made of standardized tests: the Carnegie-Mellon study in Pennsylvania and Madaus' study in Ireland. Results were similar: administrators did not use test results because they had no effect on policy. However, it is claimed that whenever test results become a key element in decisions affecting individual life chances (e.g., grade promotion), the administering agency assumes a great deal of power over the schooling process, and administrators, teachers and pupils modify their behaviour and attitudes.

McLean, L. "Educational Assessment in the Canadian Provinces". In Assessing Educational Achievement. Special issue of Educational Analysis 4 (1982), pp. 79-96.

This article compares the methodologies and unresolved issues of analysis, interpretation and reporting, and follow-up to assessments. In the section on unresolved issues, item pools, constructed responses, unit of analysis, curriculum domain and validity are mentioned. In closing, the author stresses the importance of making assessment results available at and relevant to different levels in the educational process, particularly at the classroom level.

O'Donnell, D.H. "Assessment Within Schools: A Study of One County". Educational Research 24 (1981), pp. 43-48.

Before starting an assessment program for an entire local authority area in the U.K., a survey was conducted in one county in the area to find out what assessment procedures schools had been using to describe and monitor student performance. 317 primary, middle, and secondary schools responded to "a questionnaire setting out their use of tests, both internally and as a means of providing information on transition" (p. 43). Educators expressed the need for procedures to communicate information about children moving between schools.

Omvig, C.P. "Effects of Guidance on the Results of Standardized Achievement Testing". Measurement and Evaluation in Guidance 4 (1971), pp. 47-51.

Omvig reviewed the literature on student motivation and concluded that students tend to be careless and unmotivated in their performance on tests unless they are personally concerned about their own test scores. He was particularly interested in the problem of student motivation on standardized achievement tests used in studies of the effects of school size or class size.

This was the only article found that described an experimental study in which an attempt was made to solve the problem of student apathy towards standardized tests. A "pre-test" session was designed in which a school counsellor discussed with individual students their past standardized test results, drawing attention to those areas in which poor achievement had been displayed and praising the student for progress in an area. It was hypothesized that this "treatment" would produce more valid standardized test scores. However, the results for 270 ninth grade students did not support this hypothesis.

Power, C., and Wood, R. "National Assessment: A Review of Programs in Australia, the United Kingdom, and the United States". Comparative Education Review 28 (1984), pp. 355-377.

This is an extremely informative review that describes and compares three national programs designed to "define, assess, and monitor student achievement at a national level" (p. 355). The programs are the American National Assessment of Educational Progress (NAEP), the British Assessment of Performance Unit (APU), and the Australian Studies in Student Performance (ASSP). After relating the origin and background of each program, Power and Wood compare the programs on several counts, including political issues (e.g., teacher cooperation), technical issues (e.g., problem definition, content validity, sampling procedures and reporting), and impact areas (e.g., policy, accountability, and educational practice).

Saylor, G. "How to Use the Findings from National Assessment (NAEP)". Education Digest 40 (1974), pp. 42-45.

Ten years after the inception of the NAEP, the author discusses the lack of thought given to using the assessment results. It is reported that many evaluation specialists consider the most significant use of the NAEP to be the sets of test exercises used to measure educational attainments. For curriculum planners, Saylor felt that the NAEP test results could be used as one of a number of inputs in the preparation of curriculum materials for a particular group of students. His principal recommendation on the use of NAEP data (as available in 1974) for educational planning was that "national organizations of educators, teachers, and scholars in a subject field [ought to] establish a committee that will winnow out of the NAEP data some recommendations on how school systems, curriculum planning divisions of state departments of education, textbook writers and publishers, and producers of instructional materials may best use these findings" (p. 45).

Sebring, P.A., and Boruch, R.F. "How is National Assessment of Educational Progress Used?". Educational Measurement: Issues and Practice 2 (1983), pp. 16-20.

This is a report of an exploratory study on the uses made of NAEP results. Case studies of seven state education agencies' use of NAEP were conducted, and uses made of the results by some school districts within these states were explored. Three broad categories of use were revealed by the study: professional, policy and research. Most uses by state agencies, schools, and curriculum organizations fell into the category of professional use which referred to "employing NAEP data, methods, and materials to improve educational programs and instruction" (p. 17). Policy use occurred when NAEP data were exploited to inform decision makers at the federal or state level, or to assist federal and state agencies regulate the use of educational funds. Research use referred to the use of NAEP data to develop new measurement techniques or to understand the relationship between educational attainment and certain student and school background variables.

Sebring and Boruch point out that most previous reviews of the NAEP assumed use to have occurred only if NAEP data had been instrumental in making decisions. From their exploratory study, they claim that the use of NAEP methods and procedures are as important as the use of NAEP data. One conclusion of their report is that the NAEP does not have an efficient system for monitoring use and, without such a system, it is impossible to judge statements about high or low use (p. 20). A second conclusion is that the ambiguity of the term "use" creates a major difficulty in investigating use of the NAEP. They recommend that "use" be defined in terms of audiences (e.g., local and state agencies or federal agencies), types of use (i.e., professional, policy and research), functional nature of use (e.g., decision making or enhancement of understanding), and elements of the NAEP that are used (e.g., test items, sampling methods or data).

Shoemaker, D.M. "Applicability of Item Banking and Matrix Sampling to Educational Assessment". In Advances in Psychological and Educational Measurement, edited by D.N.M. de Gruijter and L.J.T. van der Kamp, pp. 225-231. London: Wiley, 1975.

This is one of five papers on item banking presented at the Second International Symposium on Educational Testing in 1975. From Shoemaker's point of view, item banking and matrix sampling were complimentary technologies that would place educational assessment on a firm foundation. His paper consists of three parts. First, he outlines a framework for achievement testing that includes the components of item banking and matrix sampling. Then, he discusses how the technology of item banking, and not the past goals of item banking, is of major importance to the future of achievement testing. Finally, he proposes multiple matrix sampling as a procedure for program evaluation and also for the framework of achievement testing.

Theisen, G.L.; Achola, P.P.W.; and Boakari, F.M. "The Underachievement of Cross-national Studies of Achievement". Comparative Education Review 27 (1983), pp. 46-68.

Thiesen et al. discuss the shortcomings in the design and analysis of cross-national studies that prohibit analysis of within-country factors relating to performance of achievement. One methodological limitation to conducting within-country analyses from cross-national studies is considered to be the typical sampling strategies that reflect only aggregate levels of achievement, with individual students being the unit of measurement and national systems, the unit of analysis. A second limitation has been the failure to collect data related to the social, demographic, and environmental characteristics associated with school settings. As an example, Thiesen et al. refer to the IEA Second International Mathematics Study (SIMS).

"Conspicuously missing are items dealing with school selectivity, general level of district resources, local occupational opportunities, socioeconomic status of local residents, school learning environment, or related indicators of economic/cultural context" (p. 47).

To enhance the interpretation of national data, the authors suggest seven clusters of variables be measured (p. 67).

Walberg, H.J.; Haertel, G.D.; Pascarella, E.; Junker, L.K.; and Boulanger, F.D. "Probing a Model of Educational Productivity with National Assessment Samples of Early Adolescents". American Educational Research Journal 18 (1981), pp. 233-249.

The authors attempt to test a psychological theory of educational productivity and to explore the usefulness of the NAEP data for secondary analysis for policy purposes. Science achievement scores of 2,346 13-year-old students were regressed on indices of their socio-economic status, motivation, quality (of instruction), class (social environment), and home conditions. The results indicated that class (social psychological environment) was the only index that showed a strong relationship to science achievement.

Womer, F.B., and Mastie, M.M. "Can National Assessment Change American Education?". Compact 6 (1972), pp. 26-28.

This is a discussion of the utility of NAEP results. The point is made that the NAEP was not designed to fulfill all needs of education, but rather to be just one information-gathering project, which provides general information rather than answers to specific educational questions. The authors speculate as to how the information might be used, and conclude that the ultimate success of the NAEP will depend upon teachers, administrators, board members and legislators who would use the results to improve their own decision making.

Wood, R. "Assessment Has Too Many Meanings and the One I Think We Want Isn't Clear Enough Yet". Educational Measurement: Issues and Practice 3 (1984), pp. 5-7.

Wood discusses the problem of trying to communicate the meaning of the term assessment when a number of different working definitions are in use. Two such definitions under discussion were those by Bloom (1970) and Satterly (1981) (for further information, refer to the respective author). Wood expresses the need for educational researchers to redefine the term 'assessment', and to separate it from the term 'measurement'.

__________, and Gipps, C. "An Enquiry into the Use of Test Results for Accountability Purposes". In Calling Education to Account edited by R. McCormick et al. pp. 44-54. London: Heinemann Educational Books, 1982.

This is an interim report of the U.K. Evaluation of Testing in Schools Project that was aimed at evaluating the impact of testing programs in classrooms on school practices, and on educational policies at local and national levels. One finding of the project was that anxiety about testing was much less prevalent than commonly thought, given the climate of accountability in the U.K.

__________, and Power, C. "Have National Assessments Made Us Any Wiser about 'Standards'?". *Comparative Education* 20 (1984), pp. 307-321.

This paper investigates the use of the term 'standards', "one of the most abused words in education" (p. 307). Although the authors felt that national assessment programs would contribute to a working meaning of the word, they did not see evidence of this happening yet. The effect of the 'standards' issue on the development of each of the three national assessments previously reviewed by the authors (see Power and Wood, 1984) is discussed.

B. Books, Reports and Unpublished Documents

Alberta. Minister's Advisory Committee on Student Achievement. *Student Achievement in Alberta*. Edmonton: Alberta Education, 1979.

MACOSA was established in 1976 to study the problems related to student achievement in Alberta and make recommendations for their solution. This report presents the major findings of the 18 commissioned studies, as well as MACOSA's conclusions and recommendations to Alberta Education. Of specific interest might be the chapters on (1) rationale for evaluating student achievement, (2) achievement studies conducted in language arts, mathematics, science and social studies, and (3) test development studies. Also, MACOSA's recommendations for an assessment program might be of interest.

Broadfoot, P., ed. *Selection, Certification and Control: Social Issues in Educational Assessment*. London: The Falmer Press, 1984.

This set of papers represents a comprehensive collection of issues and approaches to testing. The papers are organized under two main headings: representing theoretical perspectives and practical concerns related to policy issues. The paper by Hextall (reviewed above) was a particularly interesting critique of the APU in the United Kingdom.

Canadian Teachers' Federation. *Province-wide Student Assessment Programs - The Teachers' Response*. Winnipeg: Canadian Teachers' Federation, 1980.

A collection of papers presented at a meeting in Winnipeg, and based in part on a survey of province-wide assessment programs in Canada (Document 1 in this report is a 1980 summary of these programs). In general, the purpose of the report appears to be to assemble a wide range of personal ideas about the uses and misuses of large-scale assessment.

Canadian Teachers' Federation. Province Wide Student Assessment Programs. *A Discussion Paper*. Ottawa: Canadian Teachers' Federation, 1982.

Although the date on the title page of this report is April 1982, the date within this five-page report is February 1981. The general theme expressed is that teachers have no decision making power in establishing assessment practices in schools. The CTF discusses the problems which arise for teachers and students, and questions the benefit of testing programs on student learning. Their principles and recommendations stress that teacher organizations must begin to assess the use and implications of student assessment programs.

Gipps, C., and Goldstein, H. *Monitoring Children: An Evaluation of the Assessment of Performance Unit*. London: Heinemann Educational Books, 1983.

This is the report of an evaluation of the British Assessment of Performance Unit (APU) set up in 1974 to carry out a national monitoring program of student performance. Three chapters provide a great deal of information about the organization of the APU, with two chapters discussing the work of the three advisory bodies (the Consultative Committee, the Co-ordinating group, and the Statistics Advisory Group), and a third chapter outlining the work of the steering groups and monitoring teams. The authors conclude that the APU has had partial success: they have made progress in test development and in persuading the Local Education Authorities (LEAs) to co-operate in the assessments, but they have not had success in monitoring standards or describing changes in performance over time.

Greenbaum, W.; Garet, M.S.; and Solomon, E.R. *Measuring Educational Progress: A Study of the National Assessment*. New York: McGraw-Hill, 1977.

Part I of this book is an evaluation of the American National Assessment of Educational Progress (NAEP), or rather, an evaluation of the mid 1970's version of the NAEP. Recent developments within the NAEP have altered the program, thereby limiting the utility of this book for accurate evalation purposes. However, it is a useful historical document as it covers the NAEP's objectives and organizational development, and documents such technical procedures as the exercise development process, the sampling design and the reporting strategies.

Part II is a response of the NAEP to the foregoing evaluation. This provides an interesting rebuttal and also some insight about an assessment program in transition.

de Gruijter, D.N.M., and van der Kamp, L.J.T., eds. *Advances in Psychological and Educational Measurement*. London: Wiley, 1976.

This book contains the proceedings of the Second International Symposium on Educational Testing. The section containing 5 papers on item banking might be of interest to assessment program planners. Of particular relevance is the paper by Shoemaker, reviewed in the above section.

Nuttall, D.L., ed. *Assessing Educational Achievement*. Special Issue of Educational Analysis 4 (1982).

This collection of eleven papers has three sections dealing with (a) assessment of the individual, (b) use of the assessment of individuals as a way of assessing the performance of educational institutions or the educational system as a whole, and (c) problematic conceptual issues in the field of educational assessment. The paper by McLean on Canadian assessment (cited above) may be relevant to the present project.

Nyberg, V.R., and Lee, B. *Evaluating Academic Achievement in the Last Three Years of Secondary School in Canada*. Toronto: Canadian Education Association, 1978.

The report of a study conducted in the mid 1970's by the Canadian Education Association Committee on Evaluation and Examination Practices. A survey of all Canadian departments of education provided information on province-wide examinations; this data is summarized in Table 1 of the report (p. 15-17). Another part of the study was a survey of chief education officers of school boards, from which the following conclusions were drawn about student achievement in the last three years of secondary school: (a) final examinations were widely used in arriving at final grades; (b) chief officers felt that achievement standards had risen in the sciences, were generally the same in mathematics, and had fallen in literature and language; and (c) a substantial number of officers were dissatisfied with the lack of uniform standards of achievement across schools and school systems.

Ontario Ministry of Education. Research and Evaluation Branch. *The Ontario Assessment Instrument Pool: A Curriculum-based Aid to Evaluation*. Review and Evaluation Bulletin, vol. 1, no. 1. Toronto: Ministry of Education, 1979.

This bulletin provides a rationale and background for the Ontario Ministry of Education's decision to support the development of the Ontario Assessment Instrument Pool (OAIP). The innovation was intended to contain a wide variety of assessment methods and instruments that would serve two functions: (a) assist program evaluation at the provincial and local levels, and (b) assist the evaluation of student achievement at the classroom level for both diagnostic and summative purposes (p. 5). The report contains listings of contractual research projects related to OAIP development, and lists of the technical and advisory committee members.

Passow, A.H.; Noah, H.J.; Eckstein, M.A.; and Mallea, J.R. The National Case Study: An Empirical Comparative Study of Twenty-one Educational Systems. New York: John Wiley, 1976.

This is a complex report of a comparative study of twenty-one countries using data collected in the International Association for the Evaluation of Educational Achievement (IEA) surveys. The report was intended to reconfirm the "potential of cross-national studies of schooling, based upon broad cultural, societal and educational measures" (p. 295). However, the authors concluded that the study ought to be regarded only as an interim report, and that an explanatory model of school achievement differences was not within reach. This points to the fact that, only a decade ago, empirical comparative educational analysis was in a state of infancy.

Satterly, D. Assessment in Schools. Oxford: Blackwell, 1981.

This is a comprehensive, readable text on assessment, that would be suitable for classroom teachers. It contains chapters on such topics as test construction, interpretation of scores, test reliability, and standard setting. Satterly's general definition of 'assessment' emphasizes the importance of relating the child's learning to the child's environment:

"Educational assessment is an omnibus term which includes all the processes and products which describe the nature and extent of children's learning, its degree of correspondence with the aims and objectives of teaching and its relationship with the environments which are designed to facilitate learning" (p. 2).

Wood (1984) referred to Satterly's definition in a paper on the variation in meanings for the term 'assessment'.

Wittrock, M.C., and Wiley, D.E., eds. The Evaluation of Instruction: Issues and Problems. New York: Holt, Rinehart and Winston, 1970.

This is a collection of papers presented at a symposium on problems in the evaluation of instruction. Extremely thought-provoking is Benjamin Bloom's paper on a theory of testing that includes measurement, evaluation and assessment. The comments that follow by Michael Scriven, Gene Glass and J.P. Guilford are also of interest.